Diversity and Evolution of *Short Interspersed Nuclear Elements* (SINEs) in Angiosperm and Gymnosperm Species and their Application as molecular Markers for Genotyping

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Fakultät Biologie
des Bereiches Mathematik und Naturwissenschaften
der Technischen Universität Dresden

vorgelegt von
Dipl.-Biol. Anja Kögler

Dresden 2019

1. Gutachter

Prof. Dr. Stefan Wanke

Lehrstuhl für Botanik

Forschungsgruppe Molekulare und Organismische Diversität

Institut für Botanik

Fakultät Biologie

Technische Universität Dresden

2. Gutachter

Prof. Dr. Michael Göttfert

Lehrstuhl für Molekulargenetik

Institut für Genetik

Fakultät Biologie

Technische Universität Dresden

Eingereicht am 5. September 2019

Meinen Eltern

Hans-Jürgen & Angelika
Kögler

Table of Contents

Summary

Short interspersed nuclear elements (SINEs) are small non-autonomous and heterogeneous retrotransposons, widespread in animals and plants and usually differentially propagated in related species resulting in genome-specific copy numbers.

Within the monocots, the Poaceae (sweet grasses) is the largest and economically most important plant family. The distribution of 24 Poaceae SINE (PoaS) families, five of which showing a subfamily structure, was analyzed in five important cereals (*Oryza sativa*, *Triticum aestivum*, *Hordeum vulgare*, *Sorghum bicolor*, *Zea mays*), the energy crop *Panicum virgatum* and the model grass *Brachypodium distachyon.* The comparative investigation of SINE abundance and sequence diversity within Poaceae species provides insights into their species-specific diversification and amplification. The PoaS families and subfamilies fall into two length and structural categories: simple SINEs of up to 180 bp and dimeric SINEs larger than 240 bp. Of 24 PoaS families, 20 are structurally related across species, in particular either in their 5′ or 3′ regions. Hence, reshuffling between SINEs, likely caused by nested insertions of full-lengh and truncated copies, is an important evolutionary mechanism of SINE formation. Most striking, the recently evolved homodimeric SINE family PoaS-XIV occurs exclusively in wheat (*T. aestivum*) and consists of two tandemly arranged PoaS-X.1 copies.

Exemplary for deciduous tree species, the evolutionary history of SINE populations was examined in six Salicaceae genomes (*Populus deltoides, Populus euphratica, Populus tremula, Populus tremuloides, Populus trichocarpa, Salix purpurea*). Four of eleven Salicaceae SINE (SaliS) families exhibit a subfamily organization. The SaliS families consist of two groups, differing in their phylogenetic distribution pattern, sequence similarity and 3' end structure. These groups probably emerged at different evolutionary periods of time: during the 'salicoid duplication' (~ 65 million years ago) in the *Salix-Populus* progenitor, and during the separation of the genus *Salix* (~ 45 - 65 million years ago), respectively. Similar to the PoaS families, the majority of the 20 SaliS families and subfamilies share regions of sequence similarity, providing evidence for SINE emergence by reshuffling. Furthermore, they also contain an evolutionarily young dimeric SINE family (SaliS-V), amplified only in two poplar genomes. The special feature of the Salicaceae SINEs is the contrast of the conservation of 5' start motifs across species and SINE families compared to the high variability of

3' ends within the SINE families, differing in sequence and length, presumably resulting from mutations in the poly(A) tail as a possible route for SINE elongation. Periods of increased transpositional activity promote the dissemination of novel 3' ends. Thereby, evolutionarily older motifs are displaced leading to various 3' end subpopulations within the SaliS families. Opposed to the PoaS families with a largely equal ratio of poly(A) to poly(T) tail SINEs, the SaliS families are exclusively terminated by adenine stretches.

Among retrotransposon-based markers, SINEs are highly suitable for the development of molecular markers due to their unidirectional insertion and random distribution mainly in euchromatic genome regions, together with an easy and fast detection of the heterogeneous SINE families. As a prerequisite for the development of SINE-derived inter-SINE amplified polymorphism (ISAP) markers, 13 novel Theaceae SINE families (TheaS-I - TheaS-VII, TheaS-VIII.1 and TheaS-VIII.2, TheaS-IX - TheaS-XIII) were identified in the angiosperm tree species *Camellia japonica*. Moreover, six Pinaceae SINE families (PinS-I.1 and PinS-I.2, PinS-II – PinS-VI) were detected in the gymnosperm species *Larix decidua*. Compared to the SaliS and PoaS families, structural relationships are less frequent within the TheaS families and absent in the PinS families.

The ISAP analysis revealed the genetic identity of Europe's oldest historical camellia (*C. japonica*) trees indicating their vegetative propagation from the same ancestor specimen, which was probably the first living camellia on European ground introduced to England within the 18th century. Historical sources locate the native origin of this ancestral camellia specimen either in the Chinese province Yunnan or at the Japanese Gotō Islands. Comparative ISAPs showed no accordance to the Gotō camellia sample pool and appropriate Chinese reference samples were not available. However, the initial experiments demonstrated the potential of ISAP to resolve variations among natural populations.

The ISAP application on angiosperm trees also concerned fast growing *Populus* clones grown in short rotation coppice plantations for energy production. The species-specific *P. tremula* ISAP primers might also be applied for the discrimination of hybrid poplar clones involving *P. tremuloides* genome

portions, since SINEs of these two species are highly related. However, due to lineage-specific SINE evolution during speciation, cross-species applications are generally only successful to limited extent. The analysis of poplar hybrids composed of *P. maximowiczii* with either *P. trichocarpa* or *P. nigra* based on *P. tremula* ISAP primers showed a strongly reduced resolution.

In forestry, hybrid larch (e.g. *Larix* × *eurolepis*) genotypes have to be selected from the offspring of Japanese (*Larix kaempferi*) and European larch (*Larix decidua*) crosses, as they exhibit superior growth rates compared to the parental species. Initial ISAP-based examinations of European larch genotypes provided less polymorphic banding patterns, probably resulting from general high levels of synteny and collinearities reported for gymnosperm species. Hence, the ISAP was combined with the AFLP technique to the novel marker system inter-SINE-restriction site amplified polymorphism (ISRAP). The amplicons originating from genomic regions between SINEs and *Eco*RI cleavage sites were visualized with the sensitive capillary gel electrophoresis. The ISRAP assays, based on *Eco*RI adapter primers combined with two different SINE-derived primers, resulted in a sufficient number of polymorphic peaks to distinguish the *L. decidua* genotypes investigated. Compared to ISAPs, the ISRAP approach provides the required resolution to differentiate highly similar larch genotypes.

Acknowledgement

Above all, I am very grateful to Prof. Dr. Thomas Schmidt for providing the topic of this thesis and for the freedom to pursue own scientific interests focusing on the comparative evolution of Short interspersed nuclear element (SINE) families in related plant genomes. During my research at the chair of *Plant Cell and Molecular Biology* of the Dresden University of Technology, Prof. Dr. Thomas Schmidt assisted me with constructive proposals and valuable discussions.

I also like to appreciate the scientific support of Dr. Torsten Wenke who contributed smart suggestions for improving experiments, talks, and research strategies, and always took some time for inspiring debates and critical feedback. Moreover, as my office mate, thanks for the friendly and relaxed working atmosphere.

I want to give heartfelt thanks to Dr. Gerhard Menzel for the critical reading of this thesis, starting with helpful advices for refining the draft and progressing into greater levels of detail by improving several passages with excellent wording skills.

For outstanding and kind assistance in the lab, I am truly thankful to Nadin Fliegner. Thanks for giving me the feeling to be always welcome, for your courteous helpfulness and your friendship.
I further express my gratitude to Ines Walter for comprehensive advisory support in FISH experiments.

And last, but by no means least, I want to acknowledge the whole staff of the group *Plant Cell and Molecular Biology* for sharing their practical experience and knowledge.

Furthermore, I want to acknowledge the funding, without which this thesis would not have been possible. In this context, I also would like to thank all extern project partners who contributed to a successful cooperation.

The financial support provided by the "Schlösser, Burgen und Gärten Sachsen gemeinnützige GmbH" (Dresden, Germany) for the elucidation of the geographical origin of the Pillnitz camellia is greatly acknowledged.
Prof. Dr. Stefan Wanke, leader of the research group *Molecular and Organismic Diversity* at the chair of Botany of the Dresden University of Technology (Dresden, Germany) is thanked for initiation of this project, supervision and management.
Matthias Riedel, curator of the camellia collection at the Landschloss Pirna-Zuschendorf (Pirna, Germany), is thanked for his advisory assistance in selecting appropriate *Camellia* accessions for comparative ISAP investigations, the cultivation of seedlings of the Pillnitz camellia and, in particular, for sharing his profund knowledge about the historical theories of origin of the Pillnitz camellia.

The German Federal Ministry of Food and Agriculture (BMEL) and the Agency for Renewable Resources e.V. (FNR) are thanked for their financial support of the project "Development of retrotransposon-based molecular markers for the identification of varieties, clones and accessions as a basis for breeding, management of resources and quality control for poplar and hybrid larch" (grant 22031714, acronym TreeSINE).
I like to thank all TreeSINE colleages for their commitment and constructive cooperation:
Prof. Dr. Doris Krabel and Kristin Morgenstern of the research group *Molecular Physiology of Woody Plants* at the chair of Forest Botany of the Dresden University of Technology (Tharandt, Germany) as well as Dr. Heino Wolf, Ute Tröber and Marie Brückner of the department *Forest Genetics and Forest Tree Breeding* at the Saxony State Forestry Service (Pirna, Germany).

I also like to express my gratitude for financial support from the Dresden University of Technology, providing the short-term grant "Stipendium zur Förderung von Nachwuchswissenschaftlerinnen der TU Dresden" for the funding of young scientists.

List of Publications

(I)

Reiche, B., Kögler, A., Morgenstern, K., Brückner, M., Weber, B., Heitkam, T., Tröber, U., Meyer, M., Wolf, H., Schmidt, T., Krabel, D. (2019) Application of retrotransposon-based ISAP (inter-SINE amplified polymorphism) markers for the differentiation of common poplar genotypes. Submission to ***Canadian Journal of Forest Research*** in preparation.

(II)

Kögler, A., Seibt, K. M., Heitkam, T., Morgenstern, K., Reiche, B., Brückner, M., Wolf, H., Krabel, D., Schmidt, T. (2019) Comparative analysis of short interspersed nuclear elements (SINEs) in Salicaceae species reveals 3' end diversification in many families. Resubmitted to ***The Plant Journal*** after minor revisions on 5th of September.

(III)

Kögler, A., Schmidt, T. and Wenke, T. (2017) Evolutionary modes of emergence of short interspersed nuclear element (SINE) families in grasses. ***The Plant Journal***, 92, 676–695.

(IV)

Heitkam, T., Petrasch, S., Zakrzewski, F., Kögler, A., Wenke, T., Wanke, S. and Schmidt, T. (2015) Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest *Camellia japonica*. ***Chromosome Research***, 23, 791–806.

Chapter 1

General Introduction

1.1 Short interspersed nuclear elements (SINEs) as a subclass of non-autonomous retrotransposons

Eukaryotic genomes consist mainly of repetitive DNA sequences, which occur tandemly arranged like satellites and rRNA genes or dispersed as transposable elements (TEs). With up to 80 % and more, they are especially amplified in land plants (Feschotte *et al.*, 2002; Baucom *et al.*, 2009; Oliver *et al.*, 2013; Pellicer *et al.*, 2018). TEs promote the expansion of genome sizes, together with whole genome duplications or polyploidization (Vicient and Casacuberta, 2017; Kim, 2017). Similar to other stress conditions, polyploidization often triggers massive TE proliferation, presumably due to the temporary loss of epigenetic silencing (Slotkin and Martienssen, 2007; Parisod and Senerchia, 2012). Lineage-specific TE activity generates genetic variability and thus, contributes to genome evolution and speciation (Hua-van *et al.*, 2011; Lisch, 2013; Mascagni *et al.*, 2017). Due to their replicative propagation ('copy-and-paste'), retrotransposons are particularly invasive and induce remarkable genome size variations among species and varieties (Neumann *et al.*, 2006; Hawkins *et al.*, 2006; Gómez-Orte *et al.*, 2013).

Short interspersed nuclear elements (SINEs) are short (83 bp - 352 bp) (Deragon and Zhang, 2006; Wenke *et al.*, 2011), non-coding non-LTR retrotransposons that are propagated by the enzyme reverse transcriptase (RT), encoded by the autonomous corresponding long-interspersed nuclear elements (LINEs). In plants, only tRNA-derived SINEs were detected so far (Table 1). They are transcribed by RNA polymerase III, mediated by the internal promoter motifs, box A and box B (Galli *et al.*, 1981), which are located in the SINE 5' region. The LINE RT recognizes the SINE transcripts by their 3' tail, mostly a poly(A) or poly(T) (Dewannieux and Heidmann, 2005; Tsuchimoto *et al.*, 2008), and integrates new copies into the genome by target-primed reverse transcription (Luan *et al.*, 1993; Ostertag and Kazazian, 2001). As a result, the inherent parts of each SINE comprising 5' region, 3' region and 3' tail, are framed by unique target site duplications (TSDs) (Figure 1). The 3' tail often constitutes the only common structural feature between SINEs and LINEs (Boeke, 1997; Roy-Engel,

2012). However, they sometimes share a short region of sequence homology at the 3' end (Okada and Hamada, 1997; Baucom *et al.*, 2009; Wenke *et al.*, 2011).

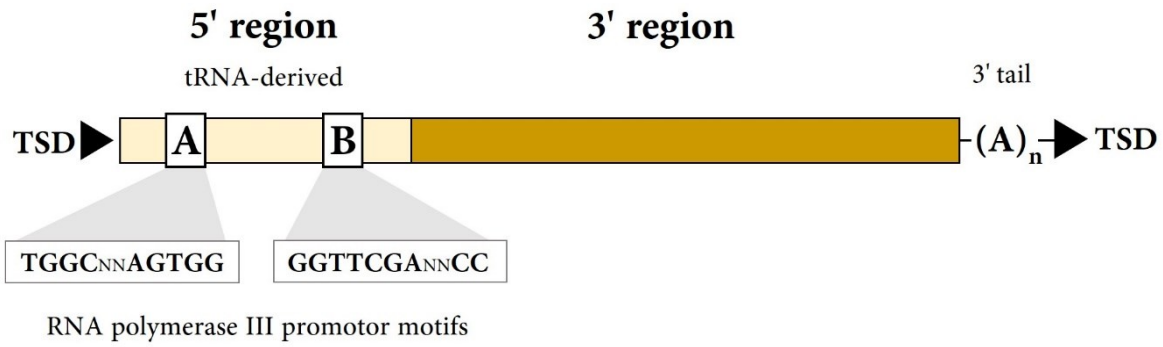


Figure 1. Typical structure of a tRNA-derived plant SINE. The conserved length of a SINE family consists of the tRNA-derived 5' region and the 3' region of unknown genomic origin. The 5' region contains the RNA polymerase III promotor, consisting of the box A and box B motif. The conserved nucleotides of the two 11 bp boxes from tRNA genes are taken from Galli *et al.* (1981). The 5' region ends 14 nucleotides after box B (Deragon and Zhang, 2006) and is relatively constant in size, contrary to the more variable 3' region. The SINE ends with the 3' tail, mostly composed of an adenine stretch of variable length, the poly(A) tail. Each SINE copy is flanked by target site duplications (TSDs), short direct repeats, resulting from the integration in the genome.

1.2 Identification and classification of the repetitive genome fraction

The *RepeatExplorer* pipeline (Novák *et al.*, 2010) enables the genome-wide detection of the major repeat families based on a graph-based clustering of next generation sequencing (NGS) reads, e.g. 454 shotgun or Illumina, covering approximately 0.02 % to 5.00 % of the respective genome size (http://repeatexplorer.org > documentation > reproducibility). However, the underlying algorithm operates inefficiently in case of low abundant and highly heterogeneous repeats such as helitrons and non-autonomous derivates of retrotransposons, e.g. SINEs, TRIMs, and transposons, e.g. MITEs (Novák *et al.*, 2010). Thus, more specialized repeat identification tools have to complement the *RepeatExplorer* analysis in order to detect the whole range of different repeat classes in genomic sequences (reviewed in Lerat, 2010). Early strategies used a homology-based search in repeat databases, e.g. *Repbase* (Jurka *et al.*, 2005) and *RepeatMasker* (Smit *et al.*, 1996-2010). However, reliable results strongly require the correct classification of the database entries and the detection of novel repeat families is excluded for highly heterogeneous repeat classes. Structure-based approaches facilitate the targeted and comprehensive identification for a certain repeat class. They are based on

conserved motifs like encoded open reading frames (ORFs) and structural features like terminal inverted repeats (TIRs) for transposons, long terminal repeats (LTRs), in addition the primer binding sites (PBS) and the polypurine tract (PPT), for LTR retrotransposons, or target site duplications (TSDs) and the poly(A) tail for non-LTR retrotransposons. Moreover, the typical size range of the element and the characteristic distances between the conserved motifs and structural features is used for their detection (reviewed in Lerat, 2010). Perspectively, machine learning-based methods might unite the detection of all repeat classes in a single pipeline (Abrusán *et al.*, 2009; Girgis, 2015; Schietgat *et al.*, 2018).

The bioinformatic tool *SINE-Finder* (Wenke *et al.*, 2011) enables the extraction of tRNA-derived SINEs from genomic sequences. A schematic representation of the underlying *Python* script is shown in Figure 2.

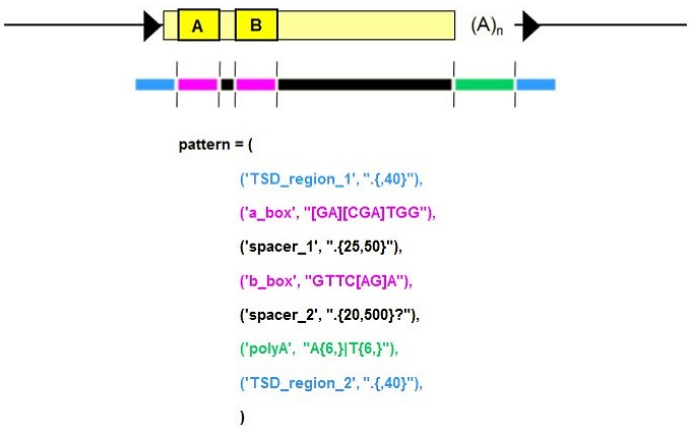


Figure 2. Principle of the *SINE-Finder* search algorithm. The SINE detection is based on the weakly conserved internal RNA polymerase III promotor (box A and B motif), the 3' tail, consisting of adenine or thymine stretches, and the target site duplications (TSDs). The distances between these features are involved in the search algorithm. The scheme was provided by Dr. Torsten Wenke.

Initially, the algorithm screens the input sequences, which might be genome assemblies, contigs, or long sequencing reads, for the weakly conserved promoter motifs of tRNA-derived SINEs. The box A motif of plant tRNA promoters TGGCNNAGTGG (Galli *et al.*, 1981) is reduced to the degenerated consensus motif 'RVTGG'among tRNA-derived plant SINEs (Wenke *et al.*, 2011). Similarily, at a distance of 25 to 50 nucleotides downstream of box A, the conserved nucleotides 'GTTCRA' within the box B motif GGTTCGANNCC (Galli *et al.*, 1981) have to be detected. In case of box A and box B

detection, the *SINE-Finder* continues 20 to 500 nucleotides downstream of box B to search for a poly(A) or poly(T) stretch, respectively, of at least six nucleotides. Finally, sequences complying these conditions are checked for TSDs in a range of 40 nucleotides preceding the box A motif and 40 nucleotides following the 3' tail. The detection of direct repeats of at least five consecutive nucleotides confirms the presence of a complete SINE copy.

The classification in biological systems, for example determination of species borders, is an anthropogenic concept bearing several problems. From an evolutionary perspective, species borders do not exist, since every 'species' is a transitional form to another species (Darwin, 1859; reviewed in Hoffmann and Blows, 1994 and Shapiro *et al.*, 2016). As a result of adaption, populations within a species evolve through 'insensibly fine gradations' (Darwin, 1859), measured in more and more ramified categories, such as subspecies, varieties, and ecotypes.

Comparable issues arise in attempts to classify the vast amount of repetitive DNA sequences within eukaryotic genomes, as repeats are also gradually evolving over time. Based on the two major transposable element (TE) classes, retrotransposons (class I) and DNA transposons (class II) (Finnegan, 1989), the categories subclass, order, clade or superfamily are defined according to typical structural features (e.g. order of ORFs, presence or absence of conserved sequence motifs) and the phylogeny of the key enzyme (e.g. reverse transcriptase or transposase protein domain sequences) (Jurka *et al.*, 2005; Wicker *et al.*, 2007; Kapitonov and Jurka, 2008; Kapitonov *et al.*, 2009; reviewed in Piégu *et al.*, 2015). The family structure is determined by DNA sequence conservation. Following the first unified hierarchical TE classification system (Wicker *et al.*, 2007), the proposed threshold of 80 % sequence similarity for family definition refers to coding regions, for example internal domains, or long terminal repeat (LTR) sequences. Some non-autonomous TE classes like SINEs and other non-coding repeats like satellite DNA possess neither of them and require adapted classification rules. For tRNA-derived SINE populations, a threshold of 60 % sequence similarity over the whole length has proven its practicability based on comprehensive data from species of various plant families (Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). It involves the differentiation between SINE families based on consensus sequences, which have to show less than 60 % sequence identity, as well as the definition of

SINE family members, which should resemble the family consensus sequence with at least 60 % sequence identity.

1.3 The history of SINE discovery in plants

In 1991, the first plant SINE family was discovered unintentionally within the scope of the comparative sequencing of the *waxy* genes in *Oryza glaberrima* and *Oryza sativa*. In the latter, two 139 bp insertions were found in an intron and in the 5' flanking region of an exon, consisting of a conserved 125 bp region with different flanking direct repeats, 14 nucleotides in length (Umeda *et al.*, 1991; Mochizuki *et al.*, 1992; Hirano *et al.*, 1994). They were designated p-SINE1 (plant SINE #1), as they show all typical features of SINEs previously described in animals (Ullu and Tschudi, 1984; Deininger and Daniels, 1986; Batzer and Deininger, 1991). However, compared to the formerly known tRNA- and 7SL RNA-derived SINEs (reviewed in Weiner *et al.*, 1986 and Okada, 1991), they end with a poly(T) stretch. Next, the tobacco (*Nicotiana tabacum*) TS SINE was discovered with an extraordinary TTG repeat at the 3' end, populating introns and flanking regions of many genes (Yoshioka *et al.*, 1993).

Among the subsequently described plant SINEs are the S1 family of *Brassica napus* (Deragon *et al.*, 1994; Lenoir *et al.*, 1997), RAthE1, RathE2 and RathE3 of *Arabidopsis thaliana* (Lenoir *et al.*, 2001; Myouga *et al.*, 2001) and the Au SINE, first detected in the grass *Aegilops umbellulata* (Yasui *et al.*, 2001). Most strikingly, due to its presence in diverse Poaceae and also Solanaceae species, Au was found to be more broadly distributed than other plant SINEs. However, exceptional high abundance was only observed in *Ae. umbellulata* and bread wheat (*Triticum aestivum*), indicating recent amplification in this plant lineage.

Successively, known SINE families were characterized in more detail, for example Au in a broad range of plant species (Fawcett *et al.*, 2006). Along with the increasing availability of DNA sequences in public databases, additional families were detected by homology searches in DDBJ, EMBL, Genbank, TIGR, and TAIR. In cultivated rice p-SINE2, p-SINE3 (Xu *et al.*, 2005) and the three OsSN families (Tsuchimoto *et al.*, 2008) were described, while eleven new BoS and SN families,

respectively, were identified in *Brassica* species (Zhang and Wessler, 2005; Deragon and Zhang, 2006).

First attempts for a *de novo* SINE detection based on weakly conserved structural features were accomplished by Baucom *et al.* (2009). Consensus sequences of the internal RNA polymerase III promotor motifs box A and box B, derived from plant SINEs, were used for homology searches in the draft reference sequence of the maize (*Zea mays*) genome (Schnable *et al.*, 2009). The results were filtered according to the presence of the TSDs and the 3' tail, resulting in the identification of Au and the maize SINE families ZmSINE1 to ZmSINE3. By a similar approach, the first Fabaceae SINE families were found in assembled genomic sequences of *Lotus japonicus* and *Medicago truncatula* (Cannon *et al.*, 2006; Sato *et al.*, 2008) containing LJ_SINE-1 to LJ_SINE-3 and MT_SINE-1 to MT_SINE-3, respectively (Gadzalski and Sakowicz, 2011).

The structural SINE features were combined in a *Python* script resulting in the bioinformatic tool *SINE-Finder*, which enabled the targeted *de novo* identification of tRNA-derived SINEs from genomic sequence data. Formerly only known from a small group of taxa, including Poaceae, Solanaceae, Brassicaceae, and Fabaceae (Table 1, A), the detection of 31 SINE families in 16 plant genomes revealed the widespread occurrence of SINEs in higher plants (Table 1, B) (Wenke *et al.*, 2011). Copy numbers are extremely variable between different SINE families and among species. Furthermore, this study discovered the chimeric origin of the tobacco TS SINE, composed of the 5' region of the Solanaceae SINE SolS-VI and the 3' end of the LINE SolRTE-I including the common poly(TTG) tail. Previously, similar 'reshuffled' SINE structures were detected for some Brassicaceae SINEs (Zhang and Wessler, 2005; Deragon and Zhang, 2006) and the OsSN families of *Oryza* (Tsuchimoto *et al.*, 2008).

The majority of SINE families is distributed among several species of a plant family (Wenke *et al.*, 2011). Others are limited to a genus like p-SINE1 and p-SINE2 in *Oryza* (Mochizuki *et al.*, 1992; Xu *et al.*, 2005) or even occur only in a single species like TS in tobacco (Wenke *et al.*, 2011).

Several studies focused on the exceptional widespread occurrence of the Au SINE (Fawcett *et al.*, 2006; Yagi *et al.*, 2011; Fawcett and Innan, 2016), which probably emerged in an ancestor of gymnosperms and angiosperms approximately 350 million years ago (mya) (Jiao *et al.*, 2011). Despite

the Au sequence conservation even among distantly related plant species, the 'patchy' phylogenetic distribution contradicts the unidirectional propagation of SINEs. The absence of SINEs in a certain lineage might result from incomplete lineage sorting during species radiation (reviewed in Ray *et al.*, 2006; Walters-Conte *et al.*, 2014; Kuritzin *et al.*, 2016; Jordan *et al.*, 2018) or from extinction caused by lacking activity and degeneration of genomic copies. Au is prone to become extinct in many species (Fawcett and Innan, 2016) containing only a few copies that persisted for example in introns of genes ('safe haven') (Schwichtenberg *et al.*, 2016).

Subsequently, the SINE analysis in the Amaranthaceae revealed the highest number of 22 different SINE families within a plant family so far (Table 1, C and Au) (Schwichtenberg *et al.*, 2016). The sugar beet (*Beta vulgaris*) SINEs exhibit an increased methylation frequency of cytosines compared with their flanking regions, most likely demonstrating the epigenetic silencing by the host. SINEs show the tendency to integrate into gene-rich regions (Deragon and Zhang, 2006; Baucom *et al.*, 2009) indicating their potential influence on gene regulation. Moreover, due to their small size, SINEs are even tolerated within genic regions like introns and untranslated regions (UTRs), thereby affecting gene and genome evolution (Seibt *et al.*, 2016). For example, these SINE integrations result in the donation of exons to genes and can lead to transduction of adjacent sequence regions (Seibt *et al.*, 2016). In wheat ~ 67 % of Au SINE insertions are associated with genes (Keidar *et al.*, 2018) and ~ 38 % with transcribed regions (Ben-David *et al.*, 2013). Intronic Au SINE copies are able to induce an irregular splicing of the respective genes, probably leading to altered protein functions (Keidar *et al.*, 2018).

Table 1. The tRNA-derived plant SINEs. The published plant SINE families are listed with the species of first identification. (a) Prior to the development of the *SINE-Finder*, SINE families were only known from Poaceae (1-10), Solanaceae (11), Brassicaceae (12-26) and Fabaceae (27-32). (b) The development of the bioinformatic tool *SINE-Finder* enabled the targeted identification of 31 novel SINE families in 16 species (33-63). (c) The highest number of different SINE families in a plant family was detected in the Amaranthaceae (64-84). Recently, 13 SINE families were identified in *Crocus sativus* (85-97; Dr. G. Menzel, unpublished results).

a

#	SINE family	Species
1	Au	*Aegilops umbellulata*
2	OsSN1	*Oryza sativa*
3	OsSN2	*Oryza sativa*
4	OsSN3	*Oryza sativa*
5	p-SINE1	*Oryza sativa*
6	p-SINE2	*Oryza sativa*
7	p-SINE3	*Oryza sativa*
8	ZmSINE1	*Zea mays*
9	ZmSINE2	*Zea mays*
10	ZmSINE3	*Zea mays*
11	TS	*Nicotianum tabacum*
12	SB1 / S1Bn	*Brassica napus*
13	SB2 / RathE1 / AtSN2	*Arabidopsis thaliana*
14	SB3 / RathE2	*Arabidopsis thaliana*
15	SB4 / RathE3 / AtSN1	*Arabidopsis thaliana*
16	SB5 / BoSB5A	*Brassica oleracea*
17	SB6 / BoSB6A	*Brassica oleracea*
18	SB7 / BoSB7A	*Brassica oleracea*
19	SB8 / BoSB8A	*Brassica oleracea*
20	SB9 / BoSB9A	*Brassica oleracea*
21	SB10 / BoSB10A	*Brassica oleracea*
22	SB11 / BoSB11	*Brassica oleracea*
23	SB12 / BoSB12	*Brassica oleracea*
24	SB13 / BoSB13	*Brassica oleracea*
25	SB14 / BoSB14A	*Brassica oleracea*
26	SB15 / BoSB15	*Brassica oleracea*
27	LJ_SINE-1	*Lotus japonicus*
28	LJ_SINE-2	*Lotus japonicus*
29	LJ_SINE-3	*Lotus japonicus*
30	MT_SINE-1	*Medicago truncatula*
31	MT_SINE-2	*Medicago truncatula*
32	MT_SINE-3	*Medicago truncatula*

b

#	SINE family	Species
33	BraS-I	*Arabidopsis lyrata*
34	CucuS-I	*Cucumis sativus*
35	CucuS-II	*Cucumis sativus*
36	EuphS-I	*Manihot esculenta*
37	FabaS-I	*Medicago truncatula*
38	FabaS-II	*Medicago truncatula*
39	FabaS-III	*Medicago truncatula*
40	FabaS-IV	*Medicago truncatula*
41	FabaS-V	*Medicago truncatula*
42	FabaS-VI	*Medicago truncatula*
43	FabaS-VII	*Medicago truncatula*
44	FabaS-VIII	*Medicago truncatula*
45	FabaS-IX	*Glycine max*
46	SaliS-I	*Populus trichocarpa*
47	SaliS-II	*Populus trichocarpa*
48	SaliS-III	*Populus trichocarpa*
49	SaliS-IV	*Populus trichocarpa*
50	SaliS-V	*Populus trichocarpa*
51	ScroS-I	*Mimulus guttatus*
52	SolS-I	*Solanum tuberosum*
53	SolS-II	*Solanum tuberosum*
54	SolS-III	*Solanum tuberosum*
55	SolS-IV	*Solanum tuberosum*
56	SolS-V	*Solanum tuberosum*
57	SolS-VI	*Solanum tuberosum*
58	SolS-VII	*Solanum tuberosum*
59	VitaS-I	*Vitis vinifera*
60	PoaS-I	*Brachypodium distachyon*
61	PoaS-II	*Brachypodium distachyon*
62	NymS-I	*Nuphar advena*
63	PinS-I	*Picea glauca, Picea sitchensis, Pinus taeda*

c

#	SINE family	Species
64	AmaS-I	*Beta vulgaris*
65	AmaS-IIa-e	*Beta vulgaris*
66	AmaS-III	*Beta vulgaris*
67	AmaS-IVa	*Beta vulgaris*
	AmaS-IVb	*Spinacia oleracea*
68	AmaS-V	*Beta vulgaris*
69	AmaS-VIa	*Beta vulgaris*
	AmaS-VIb	*Patellifolia procumbens*
70	AmaS-VII	*Beta vulgaris*
71	AmaS-VIII	*Beta vulgaris*
72	AmaS-IX	*Beta vulgaris*
73	AmaS-X	*Beta vulgaris*
74	AmaS-XI	*Beta vulgaris*
75	AmaS-XII	*Beta vulgaris*
76	AmaS-XIII	*Beta vulgaris*
77	AmaS-XIV	*Beta vulgaris*
78	AmaS-XV	*Beta vulgaris*
79	AmaS-XVI	*Beta vulgaris*
80	AmaS-XVII	*Beta vulgaris*
81	AmaS-XVIII	*Beta vulgaris*
82	AmaS-XIX	*Beta vulgaris*
83	AmaS-XX	*Beta lomatogona*
84	AmaS-XXI	*Chenopodium quinoa*

d

#	SINE family	Species
85	CroSINE-1a-c	*Crocus sativus*
86	CroSINE-2a-d	*Crocus sativus*
87	CroSINE-3a-b	*Crocus sativus*
88	CroSINE-4a-c	*Crocus sativus*
89	CroSINE-5a-b	*Crocus sativus*
90	CroSINE-6	*Crocus sativus*
91	CroSINE-7	*Crocus sativus*
92	CroSINE-8	*Crocus sativus*
93	CroSINE-9	*Crocus sativus*
94	CroSINE-10	*Crocus sativus*
95	CroSINE-11a-b	*Crocus sativus*
96	CroSINE-12	*Crocus sativus*
97	CroSINE-13	*Crocus sativus*

Only few SINE copies of a genome are able to bypass the epigenetic silencing and produce their own offspring resulting in the formation of different SINE families and subfamilies. The occurrence of SINE subfamilies was associated with the presence of several simultaneously active 'founder SINEs' per family, reflected by diagnostic nucleotide positions in the resulting subfamilies (Lenoir *et al.*, 2001). However, distinct subpopulations within a SINE family might also originate from different activity periods, in particular if a subfamily consists of evolutionarily older, more diversified copies, while the other contains younger SINEs (Yoshioka *et al.*, 1993). SINE subfamilies of comparable age structure were found for the AmaS-II family in sugar beet (Schwichtenberg *et al.*, 2016).

Stress conditions, for example drought or pathogen infestation, are known to trigger TE amplification (Negi *et al.*, 2016). However, the mechanisms allowing certain SINEs to become retrotransposition-competent, is still poorly understood. SINE amplification might be continuous over longer periods or highly increased during a short period (amplification burst) (Schwichtenberg *et al.*, 2016).

In the sweet grass family Poaceae, SINEs have been studied mainly in maize (*Zea mays*), bread wheat (*Triticum aestivum*) and domesticated rice (*Oryza sativa*) including several wild relatives, and the model grass *Brachypodium distachyon* (Table 1). The Poaceae contain many economically important cereal crops that lack SINE information, for example barley (*Hordeum vulgare*) and sorghum millet (*Sorghum bicolor*).

At the beginning of my scientific activity in 2012, many genomes like rice, maize, sorghum and *Brachypodium distachyon* (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Paterson *et al.*, 2009; Schnable *et al.*, 2009; Vogel *et al.*, 2010) were already completely sequenced or in sequencing progress like barley and wheat (Brenchley *et al.*, 2012; Mayer *et al.*, 2012; The International Wheat Genome Sequencing Consortium, 2014; Mascher *et al.*, 2017). The preliminary results of the SINE identification in the Poaceae revealed eight new SINE families in rice, wheat, sorghum, and the energy crop *Panicum virgatum* and were compiled in the diploma thesis 'Identifikation, Charakterisierung und Verbreitung von Short Interspersed Nuclear Element (SINE)-Familien in Süßgräsern (Poaceae)' (Kögler, 2012).

In this thesis, additional SINE families of barley and wheat were supplemented and all Poaceae SINE families were comparatively characterized on the molecular level, including their distribution within the Poaceae and evolutionary dynamics during species radiation (Chapter 2.2).

1.4 Molecular markers in plant breeding

Molecular markers are genomic loci showing polymorphisms between different genotypes. Their application is extremely wide-ranging and decades of research have produced numerous different marker types, more or less feasible for routine application (reviewed in: Jiang, 2013; Nybom *et al.*, 2014; Nadeem *et al.*, 2018).

Molecular markers revolutionized phylogenetic studies and became a powerful tool in plant breeding, e.g. construction of genetic linkage maps, quantitative trait locus (QTL) mapping, investigation of population diversity, germplasm analysis, cultivar genotyping and marker-assisted selection of enhanced varieties.

Early attempts to distinguish closely related individuals used the different amino acid sequence of isoenzymes (Harry, 1966; Hubby and Lewontin, 1966). The first DNA markers (restriction fragment length polymorphism - RFLP) were based on polymorphic DNA fragment lengths after cleavage with specific restriction endonucleases (Botstein *et al.*, 1980). With the development of the PCR (Mullis *et al.*, 1986; Saiki *et al.*, 1988), this ‘fingerprint’ technique was refined to amplified fragment length polymorphism (AFLP) (Vos *et al.*, 1995). Since then, a variety of PCR-based markers was developed, classified according to the type of genome (mitochondrial, chloroplast, nuclear) and source of marker development. With the exception of random amplified polymorphic DNA (RAPD) (Williams *et al.*, 1990) using short arbitrary primers, the PCR-based approaches require sequence information as prerequisite for marker development.

Simple sequence repeat (SSR) markers (Litt and Luty, 1989), also called sequence tagged microsatellite sites (STMS), are based on microsatellite polymorphisms. The highly variable number of tandemly repeated motifs constitutes an excellent source for the detection of polymorphisms and the flanking conserved genomic regions are ideal for primer design. Due to their high ubiquitous abundance and variability together with cost-effective and robust results SSRs were extensively used

over the last decade constituting the predominant marker technique in plant science, breeding and especially in population genetics (Guichoux *et al.*, 2011; Garrido-Cardenas *et al.*, 2018).

Nowadays, they are gradually replaced by sequencing-based single nucleotide polymorphism (SNP) markers like 'genotyping-by-sequencing' (GBS), or microarray-based DArT markers (Jaccoud *et al.*, 2001; Elshire *et al.*, 2011; He *et al.*, 2014). However, outside of the scientific scope they are yet less important due to extensive costs. For routine applications in practical breeding programs, the discrimination among genotypes needs to be fast, robust and cost-effective. Therefore, PCR-based markers are currently most suitable.

As different applications need adapted or mixed marker methods (Nybom, 2004), retrotransposon-derived markers (reviewed in Roy *et al.*, 2015) constitute a complementary alternative. Despite the high abundance of retrotransposons (Kumar and Bennetzen, 1999), the dispersed distribution in the genome may result in critical distances for PCR amplification, which is compensated by combination with other marker methods: outward-facing LTR-specific primers (inter-retrotransposon amplified polymorphism - IRAP) can also be combined with anchored microsatellite primers (retrotransposon-microsatellite amplified polymorphism - REMAP) (Kalendar *et al.*, 1999) or with AFLP primers (sequence-specific amplification polymorphism - S-SAP) (Waugh *et al.*, 1997).

1.5 Application of SINEs as molecular markers in plants

SINEs constitute a potential source for the development of molecular markers for phylogenetic analyses and genotyping in plant science and breeding due to several beneficial attributes.

Like other transposable elements SINEs are differentially amplified even in closely related taxa (Hawkins *et al.*, 2006; El Baidouri and Panaud, 2013; Fawcett and Innan, 2016). Furthermore, SINEs are randomly scattered throughout the genome. Although weak insertion preferences exist, for example prior to adenine stretches for poly(A) tail SINEs (Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016) or euchromatic chromosome regions due to facilitated accessibility (Schwichtenberg *et al.*, 2016), their propagation is not affected by selection. As retrotransposons are amplified replicatively, once integrated they remain inserted, allowing deduction of relationships based on presence/absence patterns (Kuritzin *et al.*, 2016). They show a tendency to form cluster (Jurka *et al.*, 2005; Seibt *et al.*,

2016), which enables the generation of PCR amplicons, as SINE copy numbers are in low to moderate range (Wenke *et al.*, 2011). SINEs are organized in highly diverse families, enabling the development of various species-specific primer combinations. Furthermore, due to their short size and the available bioinformatics tools, they can be detected easily and fast.

Polymorphic SINE insertions were used as molecular markers to resolve phylogenetic relationships of closely related species (Shedlock *et al.*, 2004). The SINE family S1 was used to elucidate relationships among wild Brassicaceae species (Tatout *et al.*, 1999) and p-SINE1 revealed the polyphyletic origin of cultivated rice (Cheng *et al.*, 2003; Ohtsubo *et al.*, 2004).

The SINE-based inter-SINE amplified polymorphism (ISAP) marker system (Seibt *et al.*, 2012; Wenke *et al.*, 2015) detects length polymorphisms of adjacent SINE copies by PCR amplification.

It is a multi-locus DNA fingerprinting method like RAPD, AFLP, and ISSR. They are considered as dominantly inherited markers, as the differentiation between heterozygotes and homozygotes based on band intensity is not feasible (Weising *et al.*, 2005). A multitude of comparative investigations was assessed and summarized in Nybom (2004), showing that RAPD, AFLP and ISSR provide similar results, although RAPDs are meanwhile outdated due to less reproducibility.

ISAPs have proven their potential for cultivar differentiation in *Solanum tuberosum* by discriminating 237 of 364 cultivars with only a single primer pair (Seibt *et al.*, 2012). Furthermore, this study revealed a resolution in the same range as observed for SSR markers. However, compared to the fast evolving microsatellite loci, ISAP markers are highly stable. They distinguished highly related parent and progeny accessions and detected somaclonal variations, emerged from *in vitro* culture (Reid and Kerr, 2007; Seibt *et al.*, 2012).

The widespread occurrence of SINEs in plants (Wenke *et al.*, 2011) opens up the possibility for an ISAP application in many different taxa.

1.6 Main objectives and outline

Depending on the specific scope a variety of molecular marker techniques can be used in plant breeding. For the selection of the appropriate marker type the concrete issue has to be agreed with marker availability for the respective taxonomic group, together with circumstances like time and costs (Nybom *et al.*, 2014). The highest resolution is achieved by SNP genotyping, recording variations between genotypes in high density genetic maps by high throughput sequencing (Ganal *et al.*, 2009).

Among PCR-based markers, SSRs have been widely applied in plant genetics (Kalia *et al.*, 2011; Mason, 2015). The universal relevance of SSR markers is based on high polymorphism rate, co-dominant inheritance, and highly stable results. Furthermore, they can be multiplexed and are easy to use. However, their development is time-consuming (Kalia *et al.*, 2011; Vieira *et al.*, 2016).

In contrast, the design of ISAP primers is cost-effective and fast, provided that assembled genome sequences are available as a basis for the *SINE-Finder*-based SINE identification.

As main objective of this thesis, the ISAP marker system was applied to angiosperm (*Camellia japonica*, *Populus tremula*) and gymnosperm *(Larix decidua*) tree species to investigate the ISAP resolution for the detection of:

- intraspecific relationships, e.g. vegetatively propagated individuals, parental genotypes and crossbred offspring (self- and cross-pollination), cultivar accessions and genetic variation within populations
- interspecific relationships, e.g. applicability of species-specific ISAP primers for genotyping in related species and discrimination of interspecific hybrids

As a prerequisite for ISAP applications, SINE families were identified in the ornamental tree *Camellia japonica*, in the European Larch *Larix decidua*, and in the Salicaceae species *Populus deltoides*, *Populus euphratica*, *Populus tremula*, *Populus tremuloides*, *Populus trichocarpa*, and *Salix purpurea*. These Salicaceae SINE (SaliS) families were analyzed concerning inter- and intraspecific divergence providing insights into their lineage-specific amplification and differential evolution to assess the transferability of species-specific ISAP primers to related species.

For an investigation of the SINE evolutionary dynamics between dicot and monocot lineages, the SaliS families and the SINE families of seven Poaceae species (*Brachypodium distachyon*, *Hordeum vulgare*, *Oryza sativa*, *Panicum virgatum*, *Sorghum bicolor*, *Triticum aestivum*, and *Zea mays*) were both analyzed regarding the following features:

- phylogenetic distribution and abundance within the plant family,
- sequence diversity, including the age of copies estimated by gradual diversification and the species-specific differentiation,
- structural relationships between the SINE families and subfamilies,
- conservation and distance between the RNA polymerase III promotor motifs, and
- chromosomal distribution.

The comparison of SINEs in the Poaceae, Salicaceae, Theaceae and Pinaceae revealed separately discussed characteristic SINE landscapes in these monocot, dicot and gymnosperm plant species.

References

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

Baidouri, M. El and Panaud, O. (2013) Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.*, **5**, 954–965.

Batzer, M.A. and Deininger, P.L. (1991) A human-specific subfamily of Alu sequences. *Genomics*, **9**, 481–487.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., SanMiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.*, **5**, e1000732.

Ben-David, S., Yaakov, B. and Kashkush, K. (2013) Genome-wide analysis of short interspersed nuclear elements SINEs revealed high sequence conservation, gene association and retrotranspositional activity in wheat. *Plant J.*, **76**, 201–210.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, **32**, 314–331.

Brenchley, R., Spannagl, M., Pfeifer, M., et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Cannon, S.B., Sterck, L., Rombauts, S., et al. (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci.*, **103**, 14959–14964.

Cheng, C., Ohtsubo, E., Ohtsubo, H., Motohashi, R., Tsuchimoto, S. and Fukuta, Y. (2003) Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. *Mol. Biol. Evol.*, **20**, 67–75.

Deininger, P.L. and Daniels, G.R. (1986) The recent evolution of mammalian repetitive DNA elements. *Trends Genet.*, **2**, 76–80.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Deragon, J.M., Landry, B.S., Pélissier, T., Tutois, S., Tourmente, S. and Picard, G. (1994) An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. *J. Mol. Evol.*, **39**, 378–386.

Dewannieux, M. and Heidmann, T. (2005) LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. *Cytogenet. Genome Res.*, **110**, 35–48.

Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379–e19379.

Fawcett, J.A. and Innan, H. (2016) High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. *Mol. Biol. Evol.*, **33**, 2593–2604.

Fawcett, J.A., Kawahara, T., Watanabe, H. and Yasui, Y. (2006) A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant Mol. Biol.*, **61**, 505–514.

Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, **3**, 329–341.

Gadzalski, M. and Sakowicz, T. (2011) Novel SINEs families in *Medicago truncatula* and *Lotus japonicus*: bioinformatic analysis. *Gene*, **480**, 21–27.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Ganal, M.W., Altmann, T. and Röder, M.S. (2009) SNP identification in crop plants. *Curr. Opin. Plant Biol.*, **12**, 211–217.

Garrido-Cardenas, J.A., Mesa-Valle, C. and Manzano-Agugliaro, F. (2018) Trends in plant research using molecular markers. *Planta*, **247**, 543–557.

Gómez-Orte, E., Vicient, C.M. and Martínez-Izquierdo, J.A. (2013) Grande retrotransposons contain an accessory gene in the unusually long 3′-internal region that encodes a nuclear protein transcribed from its own promoter. *Plant Mol. Biol.*, **81**, 541–551.

Guichoux, E., Lagache, L., Wagner, S., et al. (2011) Current trends in microsatellite genotyping. *Mol. Ecol. Resour.*, **11**, 591–611.

Harry, H. (1966) Enzyme polymorphisms in man. *Proc. R. Soc. London. Ser. B. Biol. Sci.*, **164**, 298–310.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.*, **16**, 1252–1261.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H. and Li, Z. (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front.*

Plant Sci., **5**, 484.

Hirano, H.-Y., Mochizuki, K., Umeda, M., Ohtsubo, H., Ohtsubo, E. and Sano, Y. (1994) Retrotransposition of a plant SINE into the *wx* locus during evolution of rice. *J. Mol. Evol.*, **38**, 132–137.

Hua-van, A., Rouzic, A. Le, Boutin, T.S., Filée, J. and Capy, P. (2011) The struggle for life of the genome's selfish architects. *Biol. Direct*, **6**, 19.

Hubby, J.L. and Lewontin, R.C. (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, **54**, 577–594.

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

Jaccoud, D., Peng, K., Feinstein, D. and Kilian, A. (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.*, **29**, E25.

Jiang, G.-L. (2013) Molecular markers and marker-assisted breeding in plants. In S. B. Andersen, ed. *Plant breeding from laboratories to fields.* Rijeka: IntechOpen.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97.

Jordan, V.E., Walker, J.A., Beckstrom, T.O., et al. (2018) A computational reconstruction of *Papio* phylogeny using Alu insertion polymorphisms. *Mob. DNA*, **9**, 13.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V and Jurka, M. V (2005) Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. *Cytogenet. Genome Res.*, **110**, 117–123.

Kalendar, R., Grob, T., Regina, M., Suoniemi, A. and Schulman, A. (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theor. Appl. Genet.*, **98**, 704–711.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, **177**, 309–334.

Keidar, D., Doron, C. and Kashkush, K. (2018) Genome-wide analysis of a recently active retrotransposon, Au SINE, in wheat: content, distribution within subgenomes and chromosomes, and gene associations. *Plant Cell Rep.*, **37**, 193–208.

Kim, N.S. (2017) The genomes and transposable elements in plants: are they friends or foes? *Genes and Genomics*, **39**, 359–370.

Kögler, A. (2012) Identifikation, Charakterisierung und Verbreitung von Short Interspersed Nuclear Element (SINE)- Familien in Süßgräsern (Poaceae) [In German]. *Diploma thesis.* Dresden Univerisity of Technology, Germany.

Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.*, **33**, 479–532.

Kuritzin, A., Kischka, T., Schmitz, J. and Churakov, G. (2016) Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. *PLOS Comput. Biol.*, **12**, e1004812.

Lenoir, A., Cournoyer, B., Warwick, S., Picard, G. and Deragon, J.M. (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. *Mol. Biol. Evol.*, **14**, 934–941.

Lenoir, A., Lavie, L., Prieto, J.L., Goubely, C., Coté, J.C., Pélissier, T. and Deragon, J.M. (2001) The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **18**, 2315–2322.

Lisch, D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.*, **14**, 49–61.

Litt, M. and Luty, J.A. (1989) A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.*, **44**, 397–401.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.

Mascagni, F., Giordani, T., Ceccarelli, M., Cavallini, A. and Natali, L. (2017) Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). *BMC Genomics*, **18**, 634.

Mascher, M., Gundlach, H., Himmelbach, A., et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Mason, A.S. (2015) SSR genotyping. In J. Batley, ed. *Plant genotyping.* New York: Springer, pp. 77–89.

Mayer, K.F.X., Waugh, R., Langridge, P., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

Mochizuki, K., Umeda, M., Ohtsubo, E. and Ohtsubo, H. (1992) Characterization of a plant SINE, p-SINE1, in rice genomes. *Jpn. J. Genet.*, **67**, 155–166.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.*, **51**, 263–273.

Myouga, F., Tsuchimoto, S., Noma, K., Ohtsubo, H. and Ohtsubo, E. (2001) Identification and structural analysis of SINE elements in the *Arabidopsis thaliana* genome. *Genes Genet. Syst.*, **76**, 169–179.

Nadeem, M.A., Nawaz, M.A., Shahid, M.Q., et al. (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.*, **32**, 261–285.

Negi, P., Rai, A.N. and Suprasanna, P. (2016) Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. *Front. Plant Sci.*, **7**, 1448.

Neumann, P., Koblížková, A., Navrátilová, A. and Macas, J. (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. *Genetics*, **173**, 1047–1056.

Nybom, H. (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.*, **13**, 1143–1155.

Nybom, H., Weising, K. and Rotter, B. (2014) DNA fingerprinting in botany: past, present, future. *Investig. Genet.*, **5**, 1.

Ohtsubo, H., Cheng, C., Ohsawa, I., Tsuchimoto, S. and Ohtsubo, E. (2004) Rice retroposon p-SINE1 and origin of cultivated rice. *Breed. Sci.*, **54**, 1–11.

Okada, N. (1991) SINEs. *Curr. Opin. Genet. Dev.*, **1**, 498–504.

Oliver, K.R., McComb, J.A. and Greene, W.K. (2013) Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.*, **5**, 1886–1901.

Ostertag, E.M. and Kazazian H.H., J. (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.*, **11**, 2059–2065.

Parisod, C. and Senerchia, N. (2012) Responses of transposable elements to polyploidy. In M.-A. Grandbastien and J. M. Casacuberta, eds. *Plant transposable elements. Topics in current genetics, Vol. 24.* Springer, pp. 147–168.

Park, M., Park, J., Kim, S., et al. (2012) Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *Plant J.*, **69**, 1018–1029.

Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pellicer, J., Hidalgo, O., Dodsworth, S. and Leitch, I.J. (2018) Genome size diversity and its impact on the evolution of land plants. *Genes*, **9**, 88.

Ray, D.A., Xing, J., Salem, A.-H. and Batzer, M.A. (2006) SINEs of a nearly perfect character. *Syst. Biol.*, **55**, 928–935.

Reid, A. and Kerr, E.M. (2007) A rapid simple sequence repeat (SSR)-based identification method for potato cultivars. *Plant Genet. Resour.*, **5**, 7–13.

Roy, N.S., Choi, J.-Y., Lee, S.-I. and Kim, N.-S. (2015) Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. *Genes Genomics*, **37**, 141–151.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.

Sato, S., Nakamura, Y., Kaneko, T., et al. (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, **15**, 227–239.

Schnable, Patrick S, Ware, D., Fulton, R.S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schwichtenberg, K., Wenke, T., Zakrzewski, F., Seibt, K.M., Minoche, A., Dohm, J.C., Weisshaar, B., Himmelbauer, H. and Schmidt, T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229–244.

Seibt, K.M., Wenke, T., Muders, K., Truberg, B. and Schmidt, T. (2016) Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. *Plant J.*, **86**, 268–285.

Seibt, K.M., Wenke, T., Wollrab, C., Junghans, H., Muders, K., Dehmer, K.J., Diekmann, K. and Schmidt, T. (2012) Development and application of SINE-based markers for genotyping of potato varieties. *Theor. Appl. Genet.*, **125**, 185–196.

Shedlock, A.M., Takahashi, K. and Okada, N. (2004) SINEs of speciation: tracking lineages with retroposons. *Trends Ecol. Evol.*, **19**, 545–553.

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

Tatout, C., Warwick, S., Lenoir, A. and Deragon, J.-M. (1999) SINE insertions as clade markers for wild crucifer species. *Mol. Biol. Evol.*, **16**, 1614.

The International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, **345**, 1251788.

Tsuchimoto, S., Hirao, Y., Ohtsubo, E. and Ohtsubo, H. (2008) New SINE families from rice,

OsSN, with poly(A) at the 3' ends. *Genes Genet. Syst.*, **83**, 227–236.

Ullu, E. and Tschudi, C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312**, 171–172.

Umeda, M., Ohtsubo, H. and Ohtsubo, E. (1991) Diversification of the rice *Waxy* gene by insertion of mobile DNA elements into introns. *Jpn. J. Genet.*, **66**, 569–586.

Vicient, C.M. and Casacuberta, J.M. (2017) Impact of transposable elements on polyploid plant genomes. *Ann. Bot.*, **120**, 195–207.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.*, **39**, 312–328.

Vogel, J.P., Garvin, D.F., Mockler, T.C., et al. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Vos, P., Hogers, R., Bleeker, M., et al. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.*, **23**, 4407–4414.

Walters-Conte, K.B., Johnson, D.L.E., Johnson, W.E., O'Brien, S.J. and Pecon-Slattery, J. (2014) The dynamic proliferation of CanSINEs mirrors the complex evolution of feliforms. *BMC Evol. Biol.*, **14**, 137.

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.B.T. and Powell, W. (1997) Genetic distribution of Bare–1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol. Gen. Genet.*, **253**, 687–694.

Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.*, **55**, 631–661.

Weising, K., Nybom, H., Wolff, K. and Kahl, G. (2005) DNA fingerprinting in plants: principles, methods and applications. Boca Raton: Taylor & Francis/CRC press.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Wenke, T., Seibt, K.M., Döbel, T., Muders, K. and Schmidt, T. (2015) Inter-SINE Amplified Polymorphism (ISAP) for rapid and robust plant genotyping. In J. Batley, ed. *Plant genotyping: methods and protocols.* New York: Springer, pp. 183–192.

Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S. V (1990) DNA

polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.*, **18**, 6531–6535.

Xu, J.-H., Osawa, I., Tsuchimoto, S., Ohtsubo, E. and Ohtsubo, H. (2005) Two new SINE elements, p-SINE2 and p-SINE3, from rice. *Genes Genet. Syst.*, **80**, 161–171.

Yagi, E., Akita, T. and Kawahara, T. (2011) A novel Au SINE sequence found in a gymnosperm. *Genes Genet. Syst.*, **86**, 19–25.

Yasui, Y., Nasuda, S., Matsuoka, Y. and Kawahara, T. (2001) The Au family, a novel short interspersed element (SINE) from *Aegilops umbellulata*. *Theor. Appl. Genet.*, **102**, 463–470.

Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. and Machida, Y. (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 6562–6566.

Zhang, X. and Wessler, S.R. (2005) BoS: a large and diverse family of short interspersed elements (SINEs) in *Brassica oleracea*. *J. Mol. Evol.*, **60**, 677–687.

Chapter 2

Diversity and Evolution of SINEs in Monocot and Dicot Species

2.1 The identification of SINEs in *Camellia japonica* and the multistage concept for SINE family and subfamily classification

Introduction

Camellias are popular ornamental plants with natural populations in Southeast to East Asia comprising at least 2,000 cultivated varieties worldwide (Southern California Camellia Society, 2016). Within the 18th century ornamental camellias were introduced to commerce and spread throughout Europe (Kaempfer, 1712; Edwards, 1747; Aiton, 1789; Haikal, 2008). Three old camellia trees (*Camellia japonica L.*) remained preserved from this early period and grow in Campo Bello (Vila Nova de Gaia, Portugal), Caserta (Caserta, Italy) and Pillnitz (Dresden, Germany). Since its planting at the Pillnitz Castle Park in 1801, the 'Pillnitz camellia' annually becomes a famous tourist attraction during its flowering period between February and April.

However, the geographical origin of the Pillnitz camellia remained unclear and was subject to intense research (Booth, 1829; Kümmel, 1981; Savige, 1985; Hansen, 1999; Short, 2005a, b; Vela *et al.*, 2009). Historical sources revealed two main theories, pointing to Japan or China, respectively (Haikal, 2008; Haikal, 2010).

The origin of the Pillnitz camellia might be elucidated by comparative molecular approaches. Genome sequencing and repeat identification enable the development of repeat-based molecular markers, like ISAP markers (Seibt *et al.*, 2012; Wenke *et al.*, 2015). Comparative analyses of the Pillnitz camellia and *Camellia* samples of potential regions of origin might probably support or disprove either of the theories of geographic origin.

The *de novo* repeat identification, classification, and annotation in reference genomes requires sufficient genomic resources, previously mainly available for the economically important tea plant *Camellia sinensis* (Lin *et al.*, 2011; Shi *et al.*, 2011; Taniguchi *et al.*, 2012). The Illumina sequencing of the 'Pillnitz camellia' produced 36 Gb of genomic sequences (Heitkam *et al.*, 2015). Based on the

computational pipeline *RepeatExplorer* (Novák *et al.*, 2010, 2013), the repeat content of the 4,6 Gb *C. japonica* genome (Huang *et al.*, 2013) was estimated at 73 %. Four major satellite families and the 5S rDNA form the most abundant genomic repeats, together comprising 12.5 % of the genome (Heitkam *et al.*, 2015).

This section exemplarily describes the identification of SINEs and their classification into families and subfamilies for the *Camellia japonica* genome as a prerequisite for the establishment of a SINE-based marker system for genotype comparisons (Chapter 3.1).

Experimental procedures

DNA extraction

Leaf material from the Pillnitz camellia tree (*Camellia japonica* L.), located at the park of Pillnitz castle (Pillnitz, Germany), was lyophilized and stored at - 80 ºC until usage. Genomic DNA was extracted using the standard protocol for genomic DNA from plant samples of the 'NucleoSpin Plant II' kit (Macherey-Nagel).

DNA sequencing

Next-generation sequencing (NGS) data of *C. japonica* were obtained by the commercial service of the biotechnology company Macrogen, Inc (Seoul, South Korea). Three sequence libraries with different insert sizes (Table 1) were sequenced in paired-end mode on an Illumina HiSeq2000 resulting in a total read count of 1,203,757,966 (~ 121 Gb).

Table 1. **Characteristics of the *C. japonica* sequence libraries.**

Sequence library	**[1]**	**[2]**	**[3]** (Heitkam *et al.*, 2015)
Insert size [bp]	180	300	500
Read count	2x 222,658,941	2x 199,678,426	2x 179,541,616
Size [Gb]	~ 45	~ 40	~ 36

Following the removal of duplicates and reads of reduced quality (Phred-Score < 20), the resulting 101 bp sequencing reads were assembled to 2,871,293 contigs using *SOAPdenovo2* (Luo *et al.*, 2012) consisting of 2,808,063,860 bp (Dr. Tony Heitkam, Chair of Plant Cell and Molecular Biology, Dresden University of Technology, Dresden).

SINE identification

The assembled genomic sequences were analyzed with the *SINE-Finder* tool (Wenke *et al.*, 2011). The parameters of the *SINE-Finder* search are listed in Table 2.

Table 2. Characteristics of the *SINE-Finder* search.

Parameter	Selected Option
File	SOAP_K65_raw_reads.scafSeq_min200nt
File size	203450439 B
Score for a match in motif rep	chunkwise
Chunk size	100000 bases
Overlap	1000 bases
Minimal wordsize of TSD seed	5
TSD mismatch tolerance	2
TSD mismatch penalty	1
TSD score cutoff	5
Direction of TSD search	FR
SSR-TSD overlap	3
Max. SSR motif length	6
Max. accepted mismatches in SSR	1
Min. repetitions in SSR	4
Score for a match in motif rep	2
SSR mismatch penalty	1
SSR score cutoff	8
Max. N content	0
Type of result file	fasta
Verbose	no

The SINE candidate sequences obtained by the *SINE-Finder* were clustered with *UCLUST* (Edgar, 2010), a high-performance clustering, alignment and search algorithm for large data sets, as indicated by the following command lines:

(1) uclust --sort seqs.fasta --output seqs_sorted.fasta
(2) uclust --input seqs_sorted.fasta --uc results.uc --id 0.60
(3) uclust --uc2fasta results.uc --input seqs_sorted.fasta --output results.fasta
(4) uclust --staralign results.fasta --output aligned.fasta

The sequences are sorted by decreasing length (1). The first list entry is used as query to form the first cluster: Sequences matching the query according to an identity threshold of at least 60 % were assigned to the cluster (2). This procedure is repeated iteratively, whereby the sorted sequences are processed consecutively. Subsequently, the resulting cluster are written to fasta format (3) and separately aligned (4).

Non-SINE cluster were removed from the *UCLUST* results. The remaining SINE candidate cluster were merged and aligned with *MAFFT* (embedded in *Geneious* Pro 6.1.8 software, standard

parameters, 200 PAM, Kearse *et al.*, 2012). The SINE cluster were separated from false positive hits by verifying the presence of structural SINE features.

SINE classification

A prerequisite for the SINE family assignment is the determination of copy numbers, which guarantees a robust consensus sequence representing the entire SINE family.

For this purpose, the SINE cluster consensus sequences were used as queries for *BLAST* (Altschul *et al.*, 1990) searches using *FASTA* (ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/fasta36/) to obtain more diversified SINE copies. An E-value maximum of 0.01 was used to limit the number of output sequences. The resulting *BLAST* hits were aligned with *MUSCLE* (Edgar, 2004). For the determination of the SINE copy number, including full-length and 5' truncated copies, the following steps were used:

(1) Recognition of the 3' end of the conserved SINE region (nucleotide upstream of the poly(A) tail) and removal of 3' truncated SINE sequences.

(2) Recognition of the SINE 5' start nucleotide by conservation in more than 50 % of all *BLAST* hits. Extraction of the 5' truncated SINE sequences and removal of diversified sequences. Derivation of a consensus sequence from the remaining full-length SINE sequences.

(3) Determination of full-length SINE copies by analysis of the first six 5' nucleotides, of which at least two have to match with the consensus sequence. Full-length SINE sequences sharing less than 60 % sequence identity to the consensus sequence were discarded.

(4) Determination of the total SINE copy number consisting of full-length and 5' truncated SINEs.

The SINE family organization in *C. japonica* results from comparisons of the newly derived consensus sequences, that have to show less than 60 % accordance to represent distinct families (Wenke *et al.*, 2011). Sometimes, the classification into families is not sufficient to display the different SINE groups. If the alignment of SINE family members shows clearly distinctive clusters, the family is separated into subfamilies. Although sharing more than 60 % sequence identity by consensus comparison, these SINE subpopulations substantially differ by diagnostic nucleotide changes, different consensus lengths or indels. However, to keep the number of subfamilies manageable, the consensus sequences of

subfamilies are not allowed to exceed 85 % similarity. The conserved length of SINE families is represented by the consensus length. The average of the pairwise identities of full-length SINE copies to the consensus sequence serves as an estimate for the diversity of the SINE family.

Visualization of the SINE family and subfamily organization

Of each TheaS family, 20 full-length copies with highest similarity to the respective consensus sequence were selected to represent the SINE family in an unrooted dendrogram. All SINE sequences (without TSDs and flanking regions) were aligned with *MAFFT* (Katoh *et al.*, 2002). Subsequently, the dendrogram was constructed using *MEGA5* (Tamura *et al.*, 2011) with the neighbor-joining distance method and maximum composite likelihood nucleotide model. The branching is based on 1000 bootstrap replications.

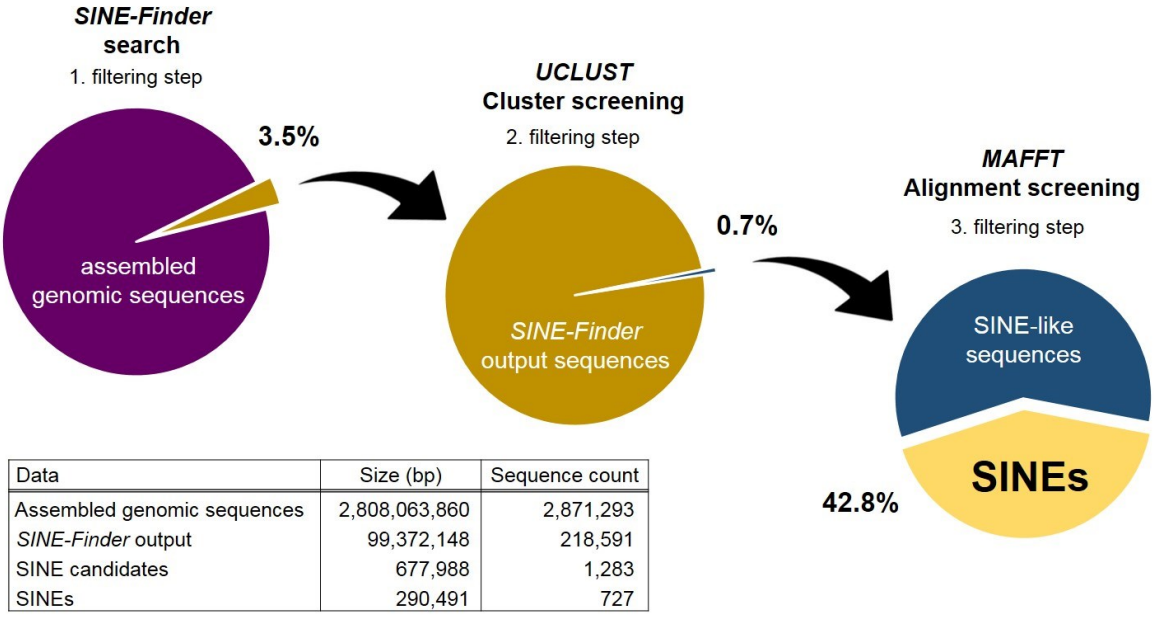
Detection of related SINE regions

Structural relationships among TheaS families were detected by an 'all-against-all' *BLAST* (*Geneious* Pro 6.1.8 software, sequence search, blastn; Kearse *et al.*, 2012) of the respective consensus sequences. Only matches with a minimum length of 30 bp (promotor region excluded) and a sequence similarity of at least 70 % were included in the analysis.

Results

SINE identification in the *Camellia japonica* genome

The SINE identification in *Camellia japonica* is based on the partially assembled genomic sequences comprising 2,871,293 contigs (Figure 1). Applying the *SINE-Finder* tool (Wenke *et al.*, 2011), SINE-like sequences were extracted and compiled to the *SINE-Finder* output, containing 3.5 % (~ 99 Mb) of the assembled genomic sequences (Figure 1). Due to weakly conserved features and high sequence heterogeneity among SINEs, the 218,591 *SINE-Finder* output sequences mainly consist of other repeats, e.g. satellite DNA.



Data	Size (bp)	Sequence count
Assembled genomic sequences	2,808,063,860	2,871,293
SINE-Finder output	99,372,148	218,591
SINE candidates	677,988	1,283
SINEs	290,491	727

Figure 1. Schematic representation of the three-step SINE identification. The *SINE-Finder* extracts all sequences from the assembled genomic sequences matching with the weakly conserved SINE features. The *SINE-Finder* results were clustered with *UCLUST* to enable the removal of false positive sequences, which are recognized by conservation over the whole sequence length. As a result, 0.7 % of the clustered *SINE-Finder* matches remained and represent SINE candidates. The SINE candidates were aligned with *MAFFT* and SINEs were separated from SINE-like sequences by the evaluation of the TSDs.

The second filtering step (*UCLUST* cluster screening) removes the bulk of false positive sequences and only 1,283 sequences of SINE candidates remained, representing 0.7 % of all *SINE-Finder* matches. The *UCLUST* output, consisting of clearly distinct sequence blocks of highly similar sequences, was screened visually. In contrast to false positives, SINE cluster are characterized by a central region of high sequence conservation, representing the SINE 5' and 3' region, which are together flanked by variable sequence regions, containing the TSDs and adjacent genomic regions.

The third filtering step (*MUSCLE* alignment screening) is based on the alignment of the 1,283 SINE candidates. They still contain 'SINE-like' false positive sequences, for example tRNA genes, among others. For the identification of SINE clusters, the presence of SINE copies within each cluster was verified by the detection of individual TSDs. Finally, the three-step identification procedure revealed 14 SINE cluster (Table 3).

Table 3. *SINE-Finder* analysis based on assembled *C. japonica* Illumina sequencing reads. The 14 SINE cluster, arranged by decreasing sequence count, are represented with the conserved consensus length and the type of the 3' tail.

SINE cluster	Consensus [bp]	Sequence count	3' tail
Cluster1	308	170	poly(A)
Cluster29	162	136	poly(T)
Cluster26	226	121	poly(A)
Cluster37	308	89	poly(A)
Cluster19	246	58	poly(A)
Cluster25	204	35	poly(A)
Cluster27	228	32	poly(A)
Cluster23	172	22	poly(A)
Cluster22	168	10	poly(A)
Cluster21	179	31	poly(A)
Cluster24	192	10	poly(A)
Cluster28	125	8	poly(A)
Cluster6	192	4	poly(A)
Cluster10	146	4	poly(A)

SINE classification into families and subfamilies

The *SINE-Finder* detects only a small fraction of SINEs matching all search criteria. Based on consensus sequences derived from these SINEs, more diversified SINE copies were identified using *BLAST* searches. After exclusion of the truncated SINE fraction and sequences showing less than 60 % sequence similarity to the SINE consensus sequence, the alignment of full-length SINE copies provides the representative SINE family consensus sequence and thereby the conserved length (Table 4). Further key characteristics, listed for each family in Table 4, are the number of SINE family members (sum of full-length and 5' truncated SINE copies) and the average sequence similarity, reflecting the intra-family diversity.

A new consensus-based comparison of the SINE cluster revealed that cluster 22 and cluster 23 together form a SINE family and represent subfamilies thereof (Table 4). Significant similarity to other

known plant SINE families could not be detected. Consequently, following the SINE designation rule of Wenke *et al.* (2011), the 13 novel SINE families were designated TheaS-I to TheaS-XIII (Theaceae SINE families).

Table 4. SINE families of the *C. japonica* genome: TheaS-I to TheaS-XIII.

SINE family	SINE cluster	Copy number			Consensus [bp]	Similarity [%]
		Full-length	5' truncated	Total		
TheaS-I	Cluster1	146	282	428	320	75
TheaS-II	Cluster29	526	802	1,328	161	81
TheaS-III	Cluster26	177	327	504	224	75
TheaS-IV	Cluster37	146	701	847	301	71
TheaS-V	Cluster19	150	142	292	246	79
TheaS-VI	Cluster25	113	148	261	204	74
TheaS-VII	Cluster27	244	226	470	224	84
TheaS-VIII.1	Cluster22	91	459	550	165	71
TheaS-VIII.2	Cluster23	34	149	183	171	85
TheaS-IX	Cluster21	41	18	59	177	87
TheaS-X	Cluster24	128	381	509	192	75
TheaS-XI	Cluster28	78	91	169	123	71
TheaS-XII	Cluster6	23	11	34	187	89
TheaS-XIII	Cluster10	274	187	461	143	72
Total		2,171	3,924	**6,095**		

The majority of TheaS families (11 of 13) range between 123 bp and 246 bp, whereas TheaS-I and TheaS-IV reached extended consensus lengths of over 300 bp (Table 4).

TheaS-II combines two special features: it represents the only SINE family containing a poly(T) tail (Table 3) and exhibits an exceptional high number of 1,328 copies (Table 4). The ratio of 5' truncated (802) to full-length copies (526) is relatively balanced compared to other extreme proportions occurring within the TheaS families: The smallest family TheaS-XII contains roughly twice as many full-length as truncated copies (23/11), while TheaS-VIII.2 mainly consists of truncated SINEs (34/149).

The similarity (Table 4) describes the diversity of a SINE family by the average of the pairwise comparisons between the SINE copies and the consensus sequence. As the diversity of copies increases with the time passed since amplification by mutations like indels and SNPs, the similarity corresponds to the average age of the copies. Therefore, evolutionarily old (~ 60 % - 70 %) and young

(~ 90 % - 100 %) SINE families are missing in the genome of *C. japonica*. With a range between 71 % and 89 % average similarity, TheaS families mostly consist of medium-aged copies. Activity profiles, representing all full-length copies of a SINE family assigned to defined similarity intervals, illustrate the age of all SINE family members (Supplementary chapter, Figure S1). Many similar, and therefore most likely evolutionarily young copies, were detected for the two SINE families TheaS-IX (34 of 50) and TheaS-XII (16 of 23) and for the subfamily TheaS-VIII.2 (18 of 34) with 90 % to 98 % similarity to the consensus sequence (Supplementary chapter, Figure S1).

To confirm the SINE classification based on percentage consensus comparisons, a dendrogram based on 20 representative full-length copies of each SINE family was constructed (Figure 2).

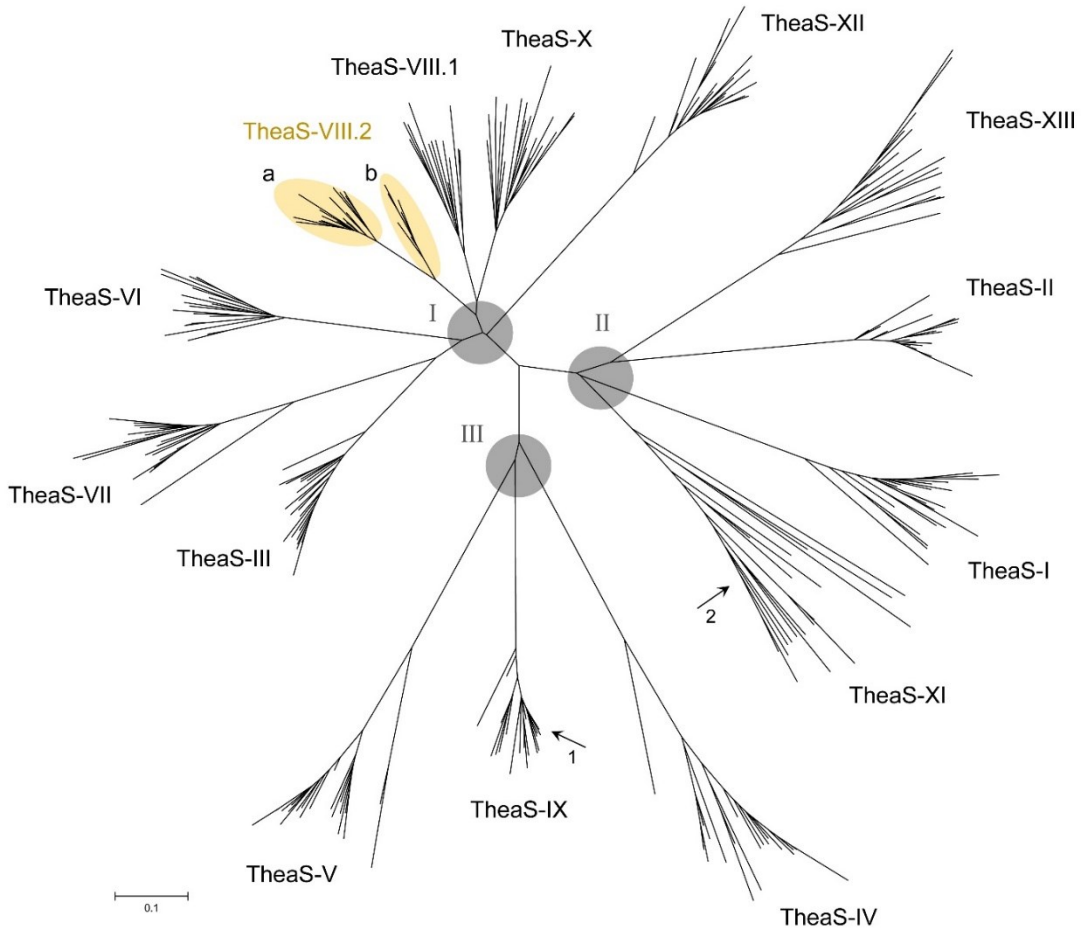
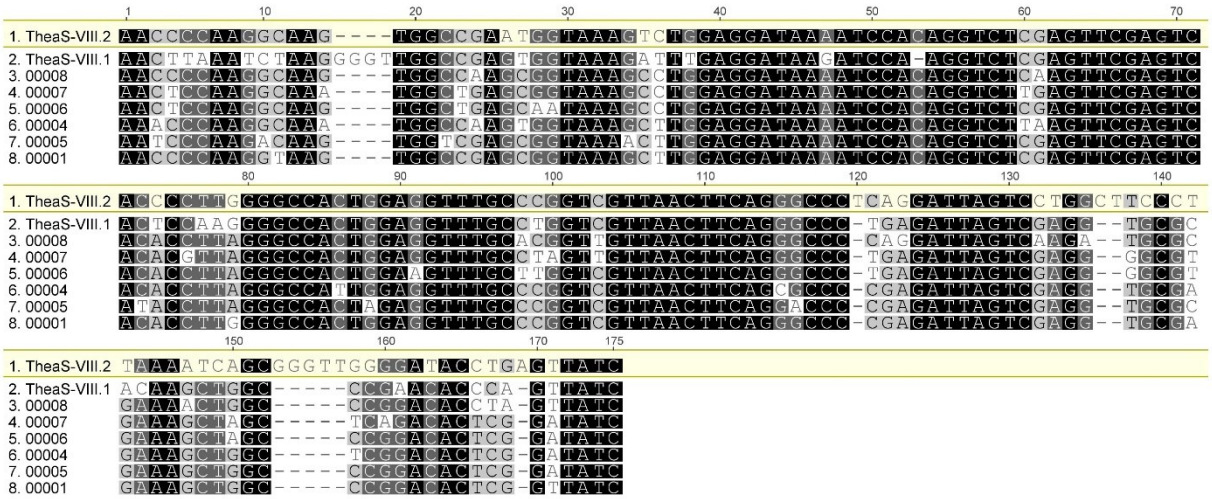


Figure 2. Dendrogram of the TheaS families. Each family is represented by 20 full-length copies with highest similarity to the respective consensus sequence. The SINE families are arranged in three main groups (I-III, grey circles). The TheaS-VIII.2 copies form two separate clades (yellow background, clade a and clade b).

The representative TheaS full-length copies were arranged to SINE family-specific clades, forming three main groups. Short branches within the clades, corresponding to highly similar copies, exemplified by TheaS-IX (Figure 2, arrow 1), are opposed to enlarged branch lengths indicating more diversified copies like observed for TheaS-XI (Figure 2, arrow 2). Moreover, six TheaS-VIII.2 copies are situated more closely to the TheaS-VIII.1 clade and obviously form an intermediate subgroup (Figure 2, yellow background, clade b).

The alignment of the six TheaS-VIII.2b copies with the TheaS-VIII.1 and TheaS-VIII.2 consensus sequences (Figure 3a) shows an insertion (Figure 3a, nucleotide position 15 – 18) and diagnostic nucleotides (Figure 3a, nucleotide position 5, 6, 9, 38, and 47, among others), which are characteristic for TheaS-VIII.2. However, they also share a diagnostic deletion typical for TheaS-VIII.1 (Figure 3a, nucleotide position 153 - 157).

a



b

Percentage identity (%)	TheaS-VIII.2	TheaS-VIII.1
TheaS-VIII.2		
TheaS-VIII.1	**73**	
00008_scaffold134834_12.4	84	78
00007_scaffold246568_4.9	76	77
00006_scaffold191535_6.6	78	79
00004_scaffold1321555_6.3	78	77
00005_scaffold444976_5.0	77	79
00001_scaffold988557_3.9	84	86

Figure 3. Comparison of the TheaS-VIII.2b copies with the TheaS-VIII.1 and TheaS-VIII.2 subfamily consensus sequences. (a) The six TheaS-VIII.2b copies with indistinct placement in the dendrogram (Figure 2) were aligned with both TheaS-VIII subfamily consensus sequences to investigate their structural relationships. (b) The pairwise similarities of the TheaS-VIII.2b copies to both TheaS-VIII subfamily consensus sequences are shown with highlightning of the highest similarity.

The pairwise similarities of the TheaS-VIII.2b copies to the TheaS-VIII.1 and TheaS-VIII.2 consensus sequences (Figure 3b) shows inconsistent results. The copy ‘00008’ shows a higher similarity to TheaS-VIII.2, while the remaining copies are similar to both subfamilies with comparable similarity values, differing only by a maximum of 2 % (SINE copy ‘00001’).

The detailed comparison of related SINE regions is a further possibility to support the SINE classification according to the 60 % similarity rule (Wenke *et al.*, 2011). Figure 4 shows structural relationships of the TheaS families obtained by consensus comparisons. Six of 13 SINE families and subfamillies (TheaS-III, TheaS-VI, TheaS-VII, TheaS-VIII.1 and TheaS-VIII.2, and TheaS-X) form a group of related SINEs.

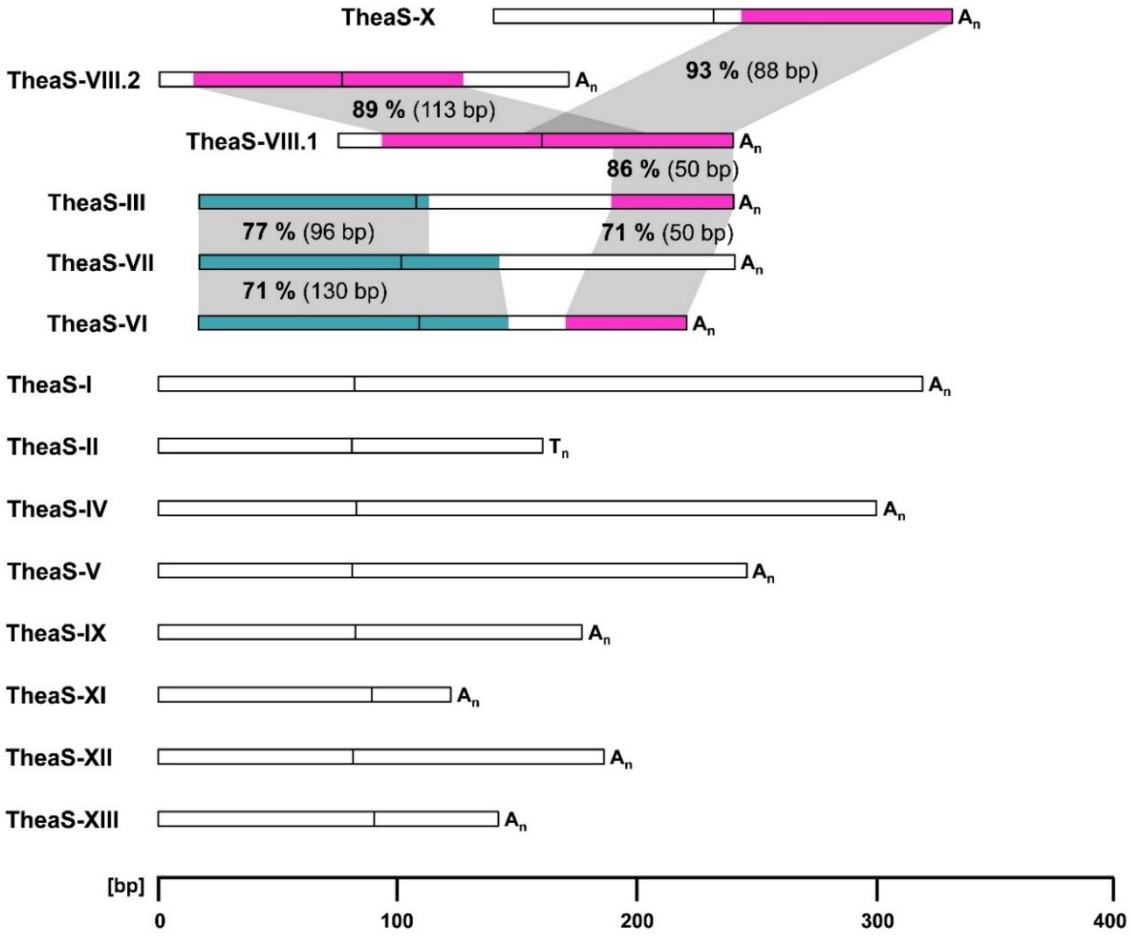


Figure 4. Structural relationships of TheaS families. The SINE families are drawn to scale and sequence regions with significant similarities are shown by the same color. The length and the similarity values of the related SINE regions are indicated by connecting grey areas. A vertical line within the schematic SINE indicates the end of the tRNA-derived 5’ region, 14 bp after the box B motif, according to Deragon and Zhang (2006).

As indicated by the visualization of the SINE classification in the dendrogram, TheaS-VIII and TheaS-X are closely related: The subfamilies TheaS-VIII.1 and TheaS-VIII.2 share a central region of 112 bp with 89 % sequence identity (Figure 4, magenta), but differ at their 5' and 3' ends. However, the overall similarity of 73 % supports their assignment to the same SINE family. TheaS-X and TheaS-VIII.1 exhibit the same 3' region, indicated by 93 % similarity over 88 bp (Figure 4, magenta), but share only 58 % over the whole length, and thus represent distinct SINE families. The 50 bp 3' end of TheaS-VIII.1 resembles the 3' end of TheaS-III and TheaS-VI, sharing 86 % and 71 % similarity, respectively. (Figure 4, magenta).

Furthermore, the group of TheaS-III, TheaS-VI, and TheaS-VII shares 5' regions of common origin (Figure 4, turquoise). TheaS-VII shows 71 % similarity over 130 bp with TheaS-VI, while similarity to TheaS-III concerns 96 bp with 77 % similarity.

Discussion

The application of the *SINE-Finder* tool combined with subsequent *BLAST* searches resulted in the identification of 6,095 SINE copies in the *Camellia japonica* genome (Table 4). These SINEs were grouped into 13 Theaceae SINE (TheaS) families based on the 60 % similarity rule (Wenke *et al.*, 2011) and the arrangement of representative copies in an unrooted dendrogram (Figure 2).

Refinement of SINE subfamily classification rules

SINE families are subject to continuous evolution. Subpopulations thereof sometimes acquire specific traits and form subfamilies.

The most intensively studied SINE subfamily structure is described for the SINE family 'Alu' of the human genome (Schmid and Deininger, 1975; Ullu and Tschudi, 1984; Willard *et al.*, 1987; Batzer and Deininger, 1991; Deininger *et al.*, 1992; Batzer *et al.*, 1996; Kapitonov and Jurka, 1996; Lander *et al.*, 2001; Teixeira-silva *et al.*, 2013). The Alu subfamilies are defined by diagnostic nucleotide positions obtained by phylogenetic analyses (reviewed in Batzer and Deininger, 2002; Deininger *et al.*, 2011). In plants, an organization into SINE subfamilies was observed in several plant families. In Solanaceae, the two subfamilies TSa and TSb were described for the TS SINE (Yoshioka *et al.*, 1993) and within the SolS families SolS-I and SolS-III are composed of two subfamilies each: SolS-Ia and SolS-Ib share 83 % consensus identity, while SolS-IIIa and SolS-IIIb share 77 %, respectively. (Wenke *et al.*, 2011). In the Brassicaceae, eight of 15 SINE families consist of subfamilies (Deragon and Zhang, 2006), while Fabaceae SINE families do not show any subfamily organization (Gadzalski and Sakowicz, 2011).

The Amaranthaceae comprise 22 SINE families and three thereof obtain subfamily populations (AmaS-IIa-e, AmaS-IVa-b, AmaS-VIa-b; Schwichtenberg *et al.*, 2016). Contrary to former studies (Yoshioka *et al.*, 1993; Deragon and Zhang, 2006; Wenke *et al.*, 2011), a group of SINEs within a family has to share 60 % to 70 % consensus similarity to form a separate subfamily (Schwichtenberg *et al.*, 2016).

Among the *C. japonica* SINEs, only TheaS-VIII exhibits a subfamily organization, detectable as distinctive clusters in the multiple sequence alignment of the SINE family members. TheaS-VIII.1 and TheaS-VIII.2 share 73 % consensus identity. Opposed to Schwichtenberg *et al.* (2016), SINE subfamilies were defined to resemble each other with 60 % to 85 % similarity by consensus comparison. SINE subpopulations sharing more than 85 % consensus similarity were not classified to subfamilies as they are too similar.

The initial family and subfamily assignment of Theaceae SINEs was largely validated by the construction of a dendrogram based on representative SINE copies (Figure 2). However, six SINEs, assigned to TheaS-VIII.2, occupy an intermediate position between the TheaS-VIII.1 clade and the main TheaS-VIII.2 clade (Figure 2, yellow background, clade a). The corresponding activity profiles (Supplementary chapter, Figure S1) suggest that TheaS-VIII.2 might have originated from TheaS-VIII.1 as it consists of evolutionarily younger copies. Thus, a diverged TheaS-VIII.1 copy might have been active, giving rise to the subgroup of TheaS-VIII.2b copies, while the TheaS-VIII.2a copies presumably originate from a later period of SINE activity. Therefore, these six copies with indistinct placement in the dendrogram might indicate the gradual differentiation of TheaS-VIII.1 to TheaS-VIII.2. Hence, according to the dendrogram topology, it might be reasonable to raise a third subfamily. However, due to the small number of copies (< 10) these SINEs are considered as a subgroup of TheaS-VIII.2.

The purpose of SINE classification is the formation of distinct groups for comparison of specific traits and the diversity between the defined groups, revealing evolutionary relationships. A dissection of subfamilies to increasingly smaller groups, for example SINE subpopulations sharing more than 85 % consensus identity or consisting of only less than ten full-length copies, might lead to impractical high numbers of subfamilies.

Furthermore, partial sequence homologies between the TheaS families (Figure 4) raise the controversial question of their classification as families or subfamilies. For example, TheaS-X, TheaS-VIII.1 and TheaS-VIII.2 share sequence regions comparable in length and similarity (Figure 4). Based on the 60 % similarity rule (Wenke *et al.*, 2011) for SINE family classification and the 60 % to 85 % range for subfamily definition, TheaS-X is classified as a separate SINE family, whereas TheaS-VIII.1

and TheaS-VIII.2 are defined as subfamilies. Contrary to the conventional subfamily definition, based on diagnostic single nucleotide polymorphisms (SNPs) (Yoshioka *et al.*, 1993; Deragon and Zhang, 2006; Wenke *et al.*, 2011), the sequence similarity-based subfamily definition (this thesis; Schwichtenberg *et al.*, 2016) is phylogenetically not supported. However, the discovery of the frequent reshuffling-based emergence of new SINEs (Chapter 2.2, Figure 7; Chapter 2.3, Figure 2) underpins the necessity of a sequence similarity-based subfamily definition.

If strictly following the conventional, SNP-based subfamily definition, SINE populations differing only by small indels had to be classified to SINE families. However, this procedure would lead to conflicts with the determined 60 % similarity rule for SINE family classification. However, a possible alternative to the sequence similarity-based rule for subfamily classification might be the relaxation of the conventional subfamily definition by allowing diagnostic indels up to a defined length, for example up to 20 % - 25 % of the total SINE length.

Summarizing, SINE subfamilies can emerge by subsequent accumulation of SNPs in active SINE copies during evolutionary timescales or by sequence reshuffling between different SINEs, for example by integration of 5' truncated copies into genomic SINEs (Chapter 2.2, Figure 10), template switching of the reverse transcriptase (Weiner, 2002; Nishihara *et al.*, 2006) or recombination (Takahashi and Okada, 2002; Deragon and Zhang, 2006; Yadav *et al.*, 2012). The resulting offspring populations of these chimeric SINEs are defined as a new family or subfamily, depending on consensus comparisons to the respective contributing SINE families and subfamilies.

Hence, the 60 % similarity rule is set as the main criterion for family assignment and subfamilies are recognized as subpopulations thereof, ranging between 60 % and 85 % sequence identity. This concept of SINE classification was exemplarily described and discussed for Theaceae SINEs in detail and applied to all SINEs identified in this thesis.

Validity of SINE copy numbers

Sequence homologies among SINEs raise difficulties in determination of the copy number, as 5' truncated copies can be assigned to two or more SINE families sharing the same 3' region. In general, the copy numbers given here most likely represent underestimations for two reasons:

(1)

A substantial amount of SINE copies might have escaped detection due to the usage of partially assembled genomic sequences for the SINE identification.

Regarding the *C. japonica* genome size of 4.6 Gb, the achieved coverage (~ 26 x) of Illumina raw reads (121 Gb, paired-end, 101 bp) was suitable to obtain assembled genomic sequences using *SOAPdenovo2* (recommended depth of coverage is 30 x; Lin *et al.*, 2011; Luo *et al.*, 2012). However, the *de novo* assembly of short sequencing reads omits the majority of the repetitive sequences of a genome, as the correct order of reads cannot be clearly reconstructed without including long-read sequencing techniques. Devices as the RS II sequencer (Pacific Biosciences) with mean read lengths of ~ 10 kb or the MinION (Oxford Nanopore) achieving read lengths of currently up to hundreds of kb might be combined to fill the gaps in order to obtain a nearly complete representation of the *C. japonica* genome sequence (Gordon *et al.*, 2016; Lu *et al.*, 2016; Jain *et al.*, 2018).y

(2)

SINE copies might remain undetected due to the specificity of the identification method.

The *SINE-Finder* only enables the identification of tRNA-derived SINEs. Even though the majority of SINEs is derived from ancestral cellular tRNAs (reviewed in Kramerov and Vassetzky, 2005), a few examples of SINE families, originated from other types of RNAs (7SL rRNA, 5S rRNA, 28S rRNA, and U1 snRNA) were reported in animals (Ullu and Tschudi, 1984; Kapitonov and Jurka, 2003; Longo *et al.*, 2015; Kojima and Jurka, 2015).

Furthermore, some families of tRNA-derived SINEs might remain undiscovered, as the *SINE-Finder* search parameters are too stringent. Only a single altered nucleotide of the minimal promoter motif (box A motif: RVTGG; box B motif: GTTCRA) is sufficient to impede detection. Moreover, even if the minimal promoter motifs are still conserved, the TSDs and the poly(A) tail of low-copy SINE families might be strongly mutated impeding recognition. As an example, the *SINE-Finder* failed to

detect the SINE family AtSB7, while all other known *Arabidopsis thaliana* SINE families were retrieved in a 'proof of concept' analysis described in Wenke *et al.* (2011).

The identification of 13 TheaS families in the genome of the Pillnitz camellia (*C. japonica*) provides a profound resource for the establishment of the ISAP marker system. The characterization of the SINE families constitutes an important prerequisite for the preliminary selection of the most suitable TheaS families for the ISAP primer design. The resulting banding pattern of the Pillnitz camellia will be compared with those of native Asian *C. japonica* genotypes in order obtain indications pointing to the factual geographic origin.

References

Abrusán, G., Grundmann, N., DeMester, L. and Makalowski, W. (2009) TEclass - a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**, 1329–1330.

Aiton, W. (1789) *Hortus Kewensis; or, a catalogue of the plants cultivated in the Royal Botanic Garden at Kew*, London: Printed for George Nicol, Bookseller to his Majesty, p. 460.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Batzer, M.A. and Deininger, P.L. (1991) A human-specific subfamily of Alu sequences. *Genomics*, **9**, 481–487.

Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.*, **3**, 370–379.

Batzer, M.A., Deininger, P.L., Hellmann-blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zigtkiewicz, E. and Zuckerkandl, E. (1996) Standardized nomenclature for Alu repeats. *J. Mol. Evol.*, **42**, 3–6.

Booth, W. (1829) History and description of the species of *Camellia* and *Thea* and of the varieties of the *Camellia japonica* that have been imported from China. *Trans Hortic Soc L.*, **7**, 519–562.

Darwin, C. (1859) Difficulties of the theory. In J. Murray, ed. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: W. Clowes and sons, pp. 171–206.

Deininger, P., Lander, E., Linton, L., et al. (2011) Alu elements: know the SINEs. *Genome Biol.*, **12**, 236.

Deininger, P.L., Batzer, M.A., Hutchison, C.A. and Edgell, M.H. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.*, **8**, 307–311.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32, 1792–1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**,

2460–2461.

Edwards, G. (1747) *A natural history of birds, Volume 2*. London: Royal College of Physicians, p. 53.

Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.*, **5**, 103–107.

Gadzalski, M. and Sakowicz, T. (2011) Novel SINEs families in *Medicago truncatula* and *Lotus japonicus*: bioinformatic analysis. *Gene*, **480**, 21–27.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Girgis, H.Z. (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, **16**, 227.

Gordon, D., Huddleston, J., Chaisson, M.J.P., et al. (2016) Long-read sequence assembly of the gorilla genome. *Science*, **352**, aae0344.

Haikal, M. (2008) *Das Geheimnis der Kamelie* [In German], Dresden: Sandstein Verlag.

Haikal, M. (2010) *Der Kamelienwald: Die Geschichte einer deutschen Gärtnerei* [In German], Dresden: Sandstein Verlag.

Hansen, W. (1999) Camellias in Germany – past and present. *Int Camellia J*, **31**, 112–117.

Heitkam, T., Petrasch, S., Zakrzewski, F., Kögler, A., Wenke, T., Wanke, S. and Schmidt, T. (2015) Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest *Camellia japonica*. *Chromosom. Res.*, **23**, 791–806.

Hoffmann, A. and Blows, M. (1994) Species borders: ecological and evolutionary perspectives. *Trends Ecol Evol*, **9**, 223–227.

Huang, H., Tong, Y., Zhang, Q. and Gao, L. (2013) Genome size variation among and within *Camellia* species by using flow cytometric analysis. *PLoS One*, **8**, e64981.

Jain, M., Koren, S., Miga, K.H., et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.

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

Kaempfer, E. (1712) *Amoenitatum Exoticarum Politico-Physico-Medicarum Fasciculi V.* Lemgo: Heinrich Wilhelm Meyer, pp. 850–852.

Kapitonov, V.V. and Jurka, J. (2003) A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.*, **20**, 694–702.

Kapitonov, V.V. and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412.

Kapitonov, V.V. and Jurka, J. (1996) The age of Alu subfamilies. *J. Mol. Evol.*, **42**, 59–65.

Kapitonov, V.V., Tempel, S. and Jurka, J. (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene*, **448**, 207–213.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., **30**, 3059–3066.

Kearse, M., Moir, R., Wilson, A., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kojima, K.K. and Jurka, J. (2015) Ancient origin of the U2 small nuclear RNA gene - Targeting non-LTR retrotransposons utopia. *PLoS One*, **10**, 1–16.

Kramerov, D.A. and Vassetzky, N.S. (2005) Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, **247**, 165–221.

Kümmel, F. (1981) The oldest camellias in the German democratic republic. *Am Camellia Yearb*, **36**, 164–175.

Lander, E.S., Heaford, A., Sheridan, A., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lerat, E. (2010) Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs. *Heredity*, **104**, 520–533.

Lin, J., Kudrna, D. and Wing, R.A. (2011) Construction, characterization, and preliminary BAC-end sequence analysis of a bacterial artificial chromosome library of the tea plant (*Camellia sinensis*). *J Biomed Biotechnol*, 2011, 476723.

Longo, M.S., Brown, J.D., Zhang, C., O'Neill, M.J. and O'Neill, R.J. (2015) Identification of a recently active mammalian SINE derived from ribosomal RNA. *Genome Biol. Evol.*, **7**, 775–788.

Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics*, **14**, 265–279.

Luo, R., Liu, B., Xie, Y., et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1, 18.

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

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

Nishihara, H., Smit, A.F.A. and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.*, **16**, 864–874.

Piégu, B., Bire, S., Arensburger, P. and Bigot, Y. (2015) A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.*, **86**, 90–109.

Savige, T. (1985) The ancient camellias of Europe. *Int Camellia J*, **17**, 80–82.

Schietgat, L., Vens, C., Cerri, R., Fischer, C.N., Costa, E., Ramon, J., Carareto, C.M.A. and Blockeel, H. (2018) A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLoS Comput. Biol.*, **14**, e1006097.

Schmid, C.W. and Deininger, P.L. (1975) Sequence organization of the human genome. *Cell*, **6**, 345–358.

Schwichtenberg, K., Wenke, T., Zakrzewski, F., Seibt, K.M., Minoche, A., Dohm, J.C., Weisshaar, B., Himmelbauer, H. and Schmidt, T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229–244.

Seibt, K.M., Wenke, T., Wollrab, C., Junghans, H., Muders, K., Dehmer, K.J., Diekmann, K. and Schmidt, T. (2012) Development and application of SINE-based markers for genotyping of potato varieties. *Theor. Appl. Genet.*, **125**, 185–196.

Shapiro, B.J., Leducq, J. and Mallet, J. (2016) What is speciation ? *PLoS Genet.*, **12**, e1005860.

Shi, C., Yang, H., Wei, C., et al. (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC*

Genomics, **12**, 131.

Short, H. (2005a) England's first camellias. *Int Camellia J*, **37**, 51–56.

Short, H. (2005b) The truth about Lord Petre's camellias. *Int Camellia J*, **37**, 56–59.

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

Southern California Camellia Society (2016) *Camellia nomenclature: twenty-Eighth Revised Edition.* B. D. King and R. C. Buggeln, eds., United States: CreateSpace Independent Publishing Platform.

Takahashi, K. and Okada, N. (2002) Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids. *Mol. Biol. Evol.*, **19**, 1303–1312.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.

Taniguchi, F., Fukuoka, H. and Tanaka, J. (2012) Expressed sequence tags from organ-specific cDNA libraries of tea (*Camellia sinensis*) and polymorphisms and transferability of EST-SSRs across *Camellia* species. *Breed. Sci.*, **62**, 186–195.

Teixeira-silva, A., Silva, R.M., Carneiro, J., Amorim, A. and Azevedo, L. (2013) The role of recombination in the origin and evolution of Alu subfamilies. *PLoS One*, **8**, e64884.

Ullu, E. and Tschudi, C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312**, 171–172.

Vela, P., Couselo, J., Salinero, C., González, M. and Sainz, M. (2009) Morpho-botanic and molecular characterization of the oldest camellia trees in Europe. *Int Camellia J*, **41**, 51–57.

Weiner, A.M. (2002) SINEs and LINEs: the art of biting the hand that feeds you. *Curr. Opin. Cell Biol.*, **14**, 343–350.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Wenke, T., Seibt, K.M., Döbel, T., Muders, K. and Schmidt, T. (2015) Inter-SINE Amplified Polymorphism (ISAP) for rapid and robust plant genotyping. In J. Batley, ed. *Plant genotyping: methods and protocols.* New York: Springer, pp. 183–192.

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

Willard, C., Nguyen, H.T. and Schmid, C.W. (1987) Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.*, **26**, 180–186.

Yadav, V.P., Mandal, P.K., Bhattacharya, A. and Bhattacharya, S. (2012) Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*. *Nat. Commun.*, **3**, 854.

Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. and Machida, Y. (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 6562–6566.

2.2 Evolutionary modes of SINE family emergence in grasses

This study has been published as:

Kögler, A., Schmidt, T. and Wenke, T. (2017) Evolutionary modes of emergence of short interspersed nuclear element (SINE) families in grasses. **Plant J.**, 92, 676–695.

The preliminary work for this study was submitted as diploma thesis to Prof. Dr. T. Schmidt (Chair of Plant Cell and Molecular Biology, Dresden University of Technology, Dresden, Germany):

Kögler, A. (2012) Identifikation, Charakterisierung und Verbreitung von Short Interspersed Nuclear Element (SINE)- Familien in Süßgräsern (Poaceae) [In German]. **Diploma thesis**, Dresden Univerisity of Technology, Germany.

Introduction

In plants the repetitive DNA fraction represents the largest part of the genome and hence determines the genome size. Due to their length and copy number, many different types of retrotransposons constitute the majority of the repetitive DNA (Lisch, 2013; Bennetzen and Wang, 2014).

However, a particular class of retrotransposons, designated short interspersed nuclear elements (SINEs) or retroposons, does not occupy large fractions of plant genomes. SINEs are widely scattered across the genome, often found close to or within other repeats, but also in coding regions (Lenoir *et al.*, 2001; Baucom *et al.*, 2009; Seibt *et al.*, 2016). SINEs exhibit extreme sequence diversity and different abundance between closely related species (Schwichtenberg *et al.*, 2016; Seibt *et al.*, 2016).

Plant SINEs are short (80 bp - 350 bp), non-coding and non-autonomous retrotransposons (Deragon and Zhang, 2006; Wenke *et al.*, 2011). Originally derived from tRNA genes, they are transcribed by RNA Polymerase III (Pol III), based on their internal Pol III promotor comprising a box A and box B motif (Galli *et al.*, 1981). SINEs are flanked by target site duplications (TSDs) resulting from their propagation by target-primed reverse transcription (Luan *et al.*, 1993; Ostertag and Kazazian, 2001) and terminated by a poly(A) stretch, poly(T) stretch or a simple sequence repeat (Yoshioka *et al.*, 1993; Yasui *et al.*, 2001; Kajikawa and Okada, 2002). The precise mechanism of SINE formation is still poorly understood, but their widespread distribution among eukaryotes together with an extreme structural diversity indicates their *de novo* emergence many times during evolution (Luchetti and Mantovani, 2013).

Since they are noncoding, the transposition of SINEs is likely mediated by the enzymatic machinery of active corresponding Long Interspersed Nuclear Elements (LINEs) (Jurka, 1997; Boeke, 1997; Kajikawa and Okada, 2002; Dewannieux *et al.*, 2003). The recognition of the SINE transcript by LINE proteins such as the reverse transcriptase (RT) is accomplished exclusively on the basis of the SINE tail. Only a few SINEs and LINEs show sequence similarities at their 3' end (Okada and Hamada, 1997; reviewed in Okada *et al.*, 1997; Baucom *et al.*, 2009; Wenke *et al.*, 2011). However, the origin of the tRNA-unrelated 3' region, highly variable in sequence and length, is still unknown for most SINEs.

The population of all SINEs in a genome represents a snap-shot of the dynamic process of emergence and amplification of SINE families, and diversification into SINE variants until final decay and extinction (Deininger and Batzer, 1995). Copies originating from the same ancestral SINE form a SINE family which is subject to diversification by accumulation of point mutations (reviewed in Kramerov and Vassetzky, 2005; Wenke *et al.*, 2011). The number of SINE families within a genome is highly variable ranging from a single SINE family in the Vitaceae up to 22 SINE families recently described in the Amaranthaceae (Deragon and Zhang, 2006; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). Diversification into subfamilies is common and results in species-specific SINE variants as observed, for example, in tobacco and some Amaranthaceae species (Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016; Seibt *et al.*, 2016).

In plants, SINE families have been reported in some eudicots (Solanaceae, Brassicaceae, Fabaceae, Salicaceae, Amaranthaceae), monocots (Poaceae), basal angiosperms (Nymphaeaceae), and in gymnosperms (Pinaceae, Gnetaceae) (Umeda *et al.*, 1991; Yasui *et al.*, 2001; Xu *et al.*, 2005; Fawcett *et al.*, 2006; Deragon and Zhang, 2006; Tsuchimoto *et al.*, 2008; Baucom *et al.*, 2009; Yagi *et al.*, 2011; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). The Poaceae, the fifth largest plant family comprising more than 11,000 grass species (Aliscioni *et al.*, 2012), include cereals such as wheat, rice, and maize, which are the staple food for the majority of the world population. Except rice and *Brachypodium distachyon*, cereal genomes are large, however, todays sequencing technologies make genome sequences accessible and the number and quality of sequenced grass genomes is constantly increasing. Despite the increasing amount of genomic data, the correct annotation of highly

heterogeneous SINEs, if performed at all, poses a substantial challenge, and detailed knowledge of SINEs is crucial for understanding their structure, origin, evolutionary diversification and conservation across species. Despite their impact on gene and genome evolution (Cordaux and Batzer, 2009; Deininger *et al.*, 2011; Schmitz, 2012; Seibt *et al.*, 2016), knowledge about the SINE dynamics, conservation and evolution is still limited. In this study, we represent a detailed molecular and cytogenetic analysis of SINEs in Poaceae. We describe 32 SINE families and subfamilies in grasses, relate transpositional activity during species radiation with SINE distribution and provide evidence for their reshuffling-based evolution summarized in a model for SINE family formation.

Experimental procedures

Computational methods

Poaceae sequence data, provided on NCBI homepage (http://www.ncbi.nlm.nih.gov), were compiled to a local database of 14.3 Gb containing 6,671,415 nucleotides. A list of the species analyzed and sequence data is provided in Table S1.

Genomes of wheat (http://www.ebi.ac.uk/ena/data/view/ERP000319) and barley (ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/etc/) were screened separately. For SINE identification the *SINE-Finder* algorithm (Wenke *et al.*, 2011) was used. Deviations from standard parameters are: Size of overlap (1,000 bp), TSD score cutoff (5 bp), and direction of TSD search (both directions). SINEs were selected based on the presence of the RNA Pol III promotor boxes A and B, a poly(A) or poly(T) stretch at the 3' end, and paired TSD sites. Resulting SINE cluster were built up from the aligned *SINE-Finder* hits and compared with known plant SINE consensus sequences. *BLAST* (Altschul *et al.*, 1990) searches using the consensus sequences of the identified SINE clusters as queries were performed to uncover diversified SINEs. The SINE family assignment is based on a 60 % sequence similarity threshold (Wenke *et al.*, 2011) for the delimitation of families by consensus comparisons and for the definition of SINE family members by pairwise comparisons to the consensus sequence. Separation into subfamilies was conducted in case of diagnostic nucleotide changes, indels, different consensus lengths and comparative consensus similarities below 85 %. The 3' tail sequences and TSDs were analyzed as follows: Tails must have a minimum length of 5 nucleotides, beginning at the conserved 3' end of the SINE copy (first up to fourth position following the 3' end); a mismatch in the tail sequence has to be followed by a minimum of three adenines for a poly(A) tail and three thymines for a poly(T) tail. Tail sequences differing from these criteria were classified as "not detectable". TSDs were recognized in case of a minimum length of five nucleotides allowing mismatches, if the TSD is further extended by at least three directly repeated nucleotides.

Statistical tests were used to detect potential correlations between the SINE features TSD length, 3' tail length, and similarity. The Shapiro-Wilk test was used for verification of normally distributed data. If the data failed the normality test, the Spearman's rank correlation was performed. Otherwise, the Pearson correlation coefficient was calculated.

The interspecific distribution of the identified SINE families was analyzed in the local databases of Poaceae genomes, which were based on the WGS (Whole Genome Shotgun) section of the NCBI homepage (ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/). For database searches consensus sequences of the PoaS families were used as queries (Table S2). Alignments and *BLAST* searches were implemented using stand-alone versions of *MUSCLE* (Edgar, 2004), *UCLUST* (Edgar, 2010), and *FASTA* (ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/fasta36/). Furthermore, *Geneious* Pro 6.1.7 (2005-2014 Biomatters Ltd.) was applied for *MAFFT* (Katoh *et al.*, 2002) alignments and *BLAST* searches to derive consensus elements (Table S2) and primers (Table S4). The number of transcripts of wheat SINE families was determined by NCBI megablast searches (Zhang *et al.*, 2000) in the transcriptome shotgun assembly of *Triticum aestivum* (NCBI taxid 4565). An artificial 3' tail sequence of nine adenines and thymines, respectively, was attached to the respective consensus sequences. Sequence similarities and dendrograms were calculated by *MEGA5* software (Tamura *et al.*, 2011), applying the neighbor-joining distance method and the maximum composite likelihood nucleotide model to the *MAFFT* alignment.

Plant material and DNA isolation

Seeds of wheat (*Triticum aestivum*, Chinese Spring, TRI 12922) and maize (*Zea mays*, maiz de gallina, ZEA 3511) were received from the Genbank of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. Plants were grown in a greenhouse under long day conditions. Genomic DNA was isolated from young leaves using the cetyltrimethyl ammonium bromide (CTAB) protocol (Saghai-Maroof *et al.*, 1984).

Fluorescent *in situ* hybridization

To prepare mitotic metaphase chromosomes, root tips from *T. aestivum* and *Z. mays* were synchronized as follows: Seedlings from *T. aestivum* were incubated in aerated ice water overnight with a 24 h recovery time, while seedlings from *Z. mays* were incubated in 2 mM 8-hydroxyquinoline for 4 h. Fixation of harvested root tips was carried out in methanol:acetic acid (3:1). The meristem of the root tips was macerated for 1 h at 37 °C in an enzyme solution containing 2.0 % (w/v) cellulase

from *Aspergillus niger* (Sigma), 4.0 % (w/v) cellulase Onozuka R 10 (Serva), 5 % (v/v) pectinase from *Aspergillus niger* (Sigma), 2.0 % cytohelicase from *Helix pomatia* (Sigma), and 0.5 % pectolyase from *Aspergillus japonicus* (Sigma) in citrate buffer (4 mM citric acid, 6 mM natrium citrate, pH 4.5). Chromosomes were spread onto pre-cleaned glass slides according to Schmidt *et al.* (1994). SINE family-specific probes, derived from the 3' SINE region (Table S3), were labeled by PCR with biotin-11-dUTP (Roche). *In situ* hybridization was carried out as described by Heslop-Harrison (1991). Chromosomes were counterstained with DAPI (4′,6′-diamidino-2-phenylindole) and mounted in antifade solution (Vectashield). Microscopy was executed with a Zeiss Axioplan2 Imaging fluorescent microscope using filters 02 (DAPI) and 15 (Cy3). Images were acquired with the Applied Spectral Imaging v. 3.3 software coupled with the high-resolution CCD camera ASI BV300-20A and optimized by Adobe Photoshop 7.0 software using only functions affecting the whole image equally.

Results

Structural characterization of Poaceae SINEs

For the targeted identification of SINEs we applied the *SINE-Finder* software (Wenke *et al.*, 2011) and *BLAST* analyses to scan a dataset of 144 Gb, containing sequence data of Poaceae genomes from public databases. In total, 11,052 SINE copies were retrieved and assigned to 32 families and subfamilies (Figure 1, Table S1, S2, S4). We found twelve novel PoaS (Poaceae SINE) families, designated PoaS-III to PoaS-XIV, identified in seven plant species: Rice (*Oryza sativa*), *Brachypodium distachyon,* wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), switchgrass (*Panicum virgatum*), sorghum millet (*Sorghum bicolor),* and maize (*Zea mays*). Importantly, our approach greatly expanded the number of copies of previously identified SINE families, mostly from rice and maize, by many thousands providing a robust basis for detailed characterization.

The accumulation of mutations successively leads to diversification among SINE copies and, hence, to subfamilies. For example, we identified 2,685 novel copies of OsSN2 (Tsuchimoto *et al.*, 2008), forming a diverged subfamily, which we have designated OsSN2.2 (Figure 1).

We also found remarkably diversified subfamilies for some PoaS families: PoaS-V is composed of two subfamilies, designated PoaS-V.1 and PoaS-V.2. They share 65 % sequence similarity by consensus comparison, but differ in their conserved lengths (145 bp and 140 bp), the type of 3' tail and species distribution, respectively. Also, three subfamilies have been identified for PoaS-X and PoaS-XI each. We found 18 SINE families terminating with a poly(T) tail and 14 families with a poly(A) stretch at their 3' end. Interestingly, PoaS-V occurs in two different variants: The subfamily PoaS-V.1 is characterized by a poly(T) tail, PoaS-V.2 by a poly(A) tail (Table 1).

The majority of the SINE families and subfamilies is between 108 and 178 nucleotides long, while nine families exhibit an extended length (e.g. PoaS-XIII, 244 bp; OsSN2.2, 283 bp; PoaS-XIV, 312 bp; PoaS-VII, 321 bp).

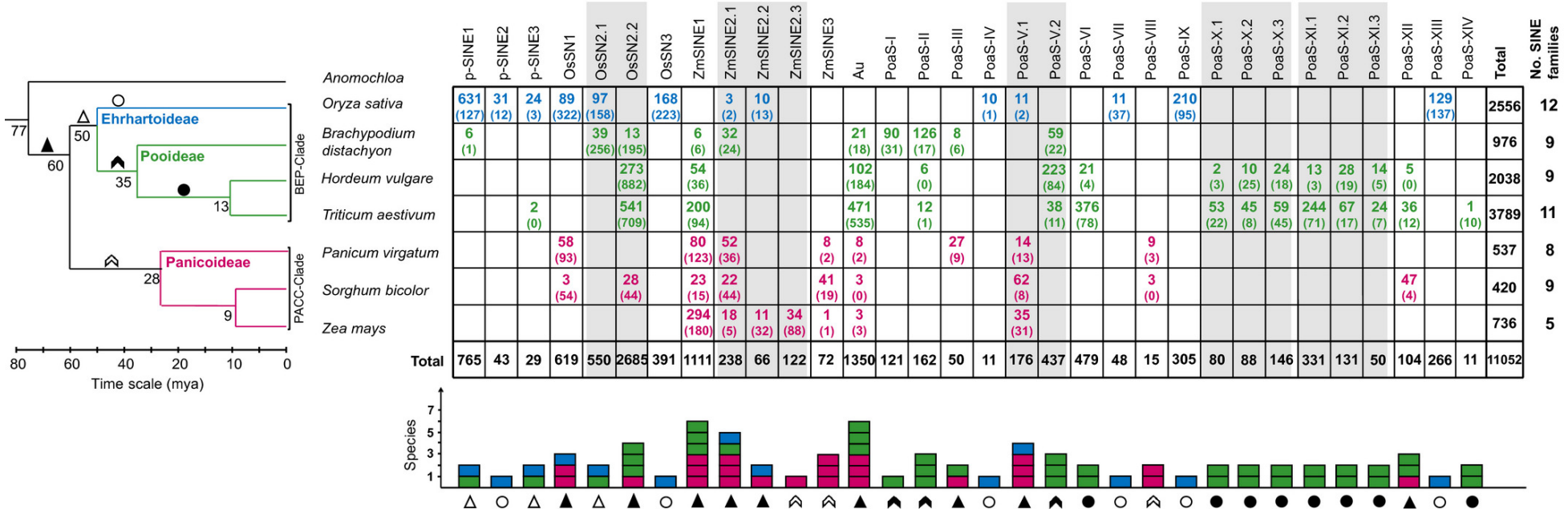


Figure 1. Phylogenetic distribution of 32 Poaceae SINE families and subfamilies. The grid shows the distribution, copy number and age of SINE families and subfamilies in Poaceae species (rows). Numbers refer to full-length SINE copies, numbers in brackets are 5′-truncated copies. Numbers (black) in bold indicate total copy numbers (full length and 5′ truncated) per SINE family (below) and per species (right). Grey shadings show SINE families with subfamily structures. Bars (color coded for Poaceae subfamilies) summarize the distribution of SINE families and subfamilies, showing the number of genomes where a SINE family or subfamily occurs. Putative periods of SINE amplification during species radiation are represented by different symbols according to the phylogenetic scheme (left). Phylogenetic relationships and divergence times are modified from Gaut (2002) and Charles *et al.* (2009).

Table 1. Structural features and abundance of Poaceae SINE families.

SINE family	Species [a]	Consensus [bp] [b]	Copy number			Similarity [%] [c]	3' tail [d] [bp]		TSD [d] [bp]
			Full-length	5' truncated	Total		Poly(A)	Poly(T)	
Au	*T. aestivum*	178	471	535	**1,006**	89		7	14
OsSN1	*O. sativa*	283	89	322	**411**	81	9		12
OsSN2.1	*O. sativa*	282	97	158	**255**	72	9		11
OsSN2.2	*T. aestivum*	283	541	709	**1,250**	80	7		10
OsSN3	*O. sativa*	176	168	223	**391**	79	9		9
p-SINE1	*O. sativa*	115	631	127	**758**	77		8	12
p-SINE2	*O. sativa*	118	31	12	**43**	74		8	12
p-SINE3	*O. sativa*	117	24	3	**27**	90		8	12
PoaS-I	*B. distachyon*	157	90	31	**121**	83	8		12
PoaS-II	*B. distachyon*	114	126	35	**161**	78	8		11
PoaS-III	*P. virgatum*	117	27	9	**36**	79	9		12
PoaS-IV	*O. sativa*	139	10	1	**11**	78		8	11
PoaS-V.1	*S. bicolor*	145	62	8	**70**	68		8	9
PoaS-V.2	*H. vulgare*	140	223	84	**307**	88	7		11
PoaS-VI	*T. aestivum*	134	376	78	**454**	88	8		12
PoaS-VII	*O. sativa*	321	11	37	**48**	78	9		12
PoaS-VIII	*P. virgatum*	108	9	3	**12**	80	10		15
PoaS-IX	*O. sativa*	128	210	95	**305**	73		8	10
PoaS-X.1	*T. aestivum*	154	53	23	**76**	97		9	16
PoaS-X.2	*T. aestivum*	152	45	8	**53**	96		8	13
PoaS-X.3	*T. aestivum*	150	59	45	**104**	74		7	10
PoaS-XI.1	*T. aestivum*	144	244	71	**315**	83		8	12
PoaS-XI.2	*T. aestivum*	146	67	17	**84**	82		7	11
PoaS-XI.3	*T. aestivum*	141	24	7	**31**	73		9	12
PoaS-XII	*S. bicolor*	144	47	4	**51**	86	8		12
PoaS-XIII	*O. sativa*	244	129	137	**266**	64	9		9
PoaS-XIV	*T. aestivum*	312	1	10	**11**	n.d.		9	16
ZmSINE1	*Z. mays*	156	294	180	**474**	75		7	10
ZmSINE2.1	*P. virgatum*	276	52	36	**88**	74		8	14
ZmSINE2.2	*Z. mays*	333	11	32	**43**	86		7	6
ZmSINE2.3	*Z. mays*	297	34	88	**122**	82		7	8
ZmSINE3	*S. bicolor*	132	41	19	**60**	88	9		14
Total			4,297	3,147	**7,444**				

[a] Species with most full-length copies.
[b] Length of consensus sequence without poly(A/T)$_n$.
[c] Average identity value of full-length copies.
[d] Average length.
n.d., not detectable (one full-length copy only).

Using the genome as a reference where most copies of a SINE family occur, we selected 4,297 full-length PoaS copies to conduct a detailed analysis of typical SINE features (Table 1).

Highly diverged SINE families are PoaS-XIII and PoaS-V.1 (64 % and 68 % average sequence identity, respectively), while highly similar copies were detected for p-SINE3 (90 %) and, in particular, for PoaS-X.2 (96 %) and PoaS-X.1 (97 %). The average similarity of SINE family members mostly ranges from 70 % to 89 % (Table S5).

By comparison of the 5' and 3' flanking regions of the 4,297 full-length copies we determined the length of the TSD enabling also the delimitation of the 3' tail length of the SINE (Table 1; Figure S1, Table S6).

We observed a positive correlation of the TSD length and the average SINE similarity (correlation factor of 0.42 and p-value of 0.01, Figure 2, Figure S2). Highly diverged SINE families such as PoaS-V.1 and PoaS-XIII have shorter TSDs than the highly conserved SINE families such as p-SINE3, PoaS-X.2 and PoaS-X.1 (Figure 2).

The TSD length reaches a maximum of 24 nucleotides for two Au copies. Average values range from 6 bp (ZmSINE2.2) to 16 bp (PoaS-X.1) (Table 1). Altogether, 616 of 4,297 (14 %) characterized full-length Poaceae SINE copies do not have a minimum TSD of 5 nucleotides (Table S6).

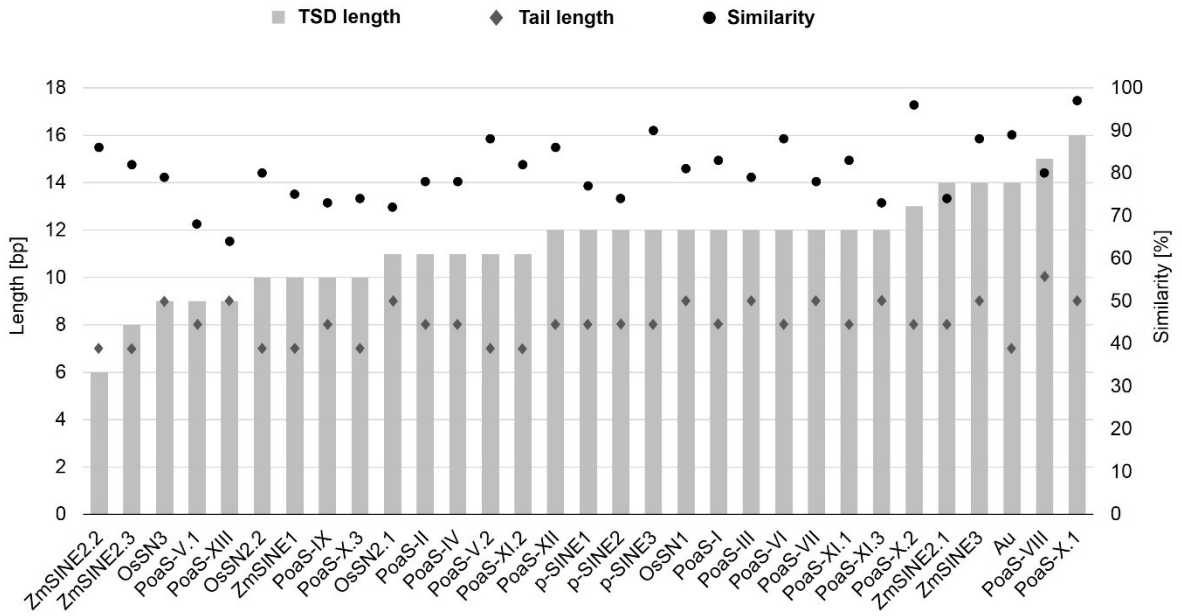


Figure 2. Relation between similarity of SINE family members, target site duplication (TSD) length and length of 3′ tails of Poaceae SINEs. The average TSD lengths (bars), arranged by increasing size, are compared with the average length of the 3′ tails (diamonds) and the average similarity of SINE family members (dots). PoaS-XIV is not included (one full-length copy only).

Averaged 3' tail lengths range between 7 bp and 10 bp (Table 1). Extreme values are 25 residues for a poly(A) and 24 residues for a poly(T) tail in individual copies of OsSN3 and p-SINE2, respectively. A 3' tail was not present in 1,032 of 4,297 (24 %) characterized full-length Poaceae SINE copies. For example, more than half of all ZmSINE1 copies (134 out of 294) in *Z. mays* do not possess a detectable poly(T) tail.

The copy number of Poaceae SINE families per genome ranges from two copies (ZmSINE3 in *Z. mays*) up to 1,250 copies of OsSN2.2 in *T. aestivum* (Figure 1, Table S3). The ratio of full-length to 5' truncated copies varies extremely between SINE families and is 3:1 in average for all SINE families investigated (Figure S3). Notably, for all three OsSN SINE families the number of 5' truncated copies exceeds, sometimes massively, the number of full-length copies (Table 1, Figure 1).

Similar to most plant SINEs, Poaceae SINEs are derived from tRNA genes and contain two sequence motifs resembling the box A and box B of the RNA polymerase III promotor. However, by comparing the 5' regions of all Poaceae SINE families with 702 Viridiplantae tRNA genes (Jühling *et al.*, 2009), no specific tRNA gene could be identified from which Poaceae SINEs may have originated (Figure S4, Table S7, Figure S5). Nevertheless, single nucleotides in box A and B are highly conserved and invariable across species and SINE families. Moreover, we found conserved 5'starts upstream of box A of the Poaceae SINEs across species. All SINE families can be assigned to one of the three typical motifs 5'-GMGAA(M)-3', 5'-GAGGA(M)-3' and 5'-GAAGGG-3' (M=A, C). However, no species-specific grouping was detected (Figure S6).

The high copy number of most SINE families prompted us to investigate the chromosomal distribution. We performed fluorescent *in situ* hybridization using a sequence shared by the PoaS-X subfamilies and a part of the 3' region of ZmSINE1 as probes on mitotic metaphase chromosomes of wheat and maize, respectively (Figure 3). Both SINE families are present on all chromosomes. In wheat, the PoaS-X subfamilies are uniformly dispersed along chromosomes up to the outermost distal regions (Figure 3a-c). In contrast, ZmSINE1 shows a moderate dispersed distribution and is largely clustered in distal ends and some centromeric regions of maize chromosomes (Figure 3d-f).

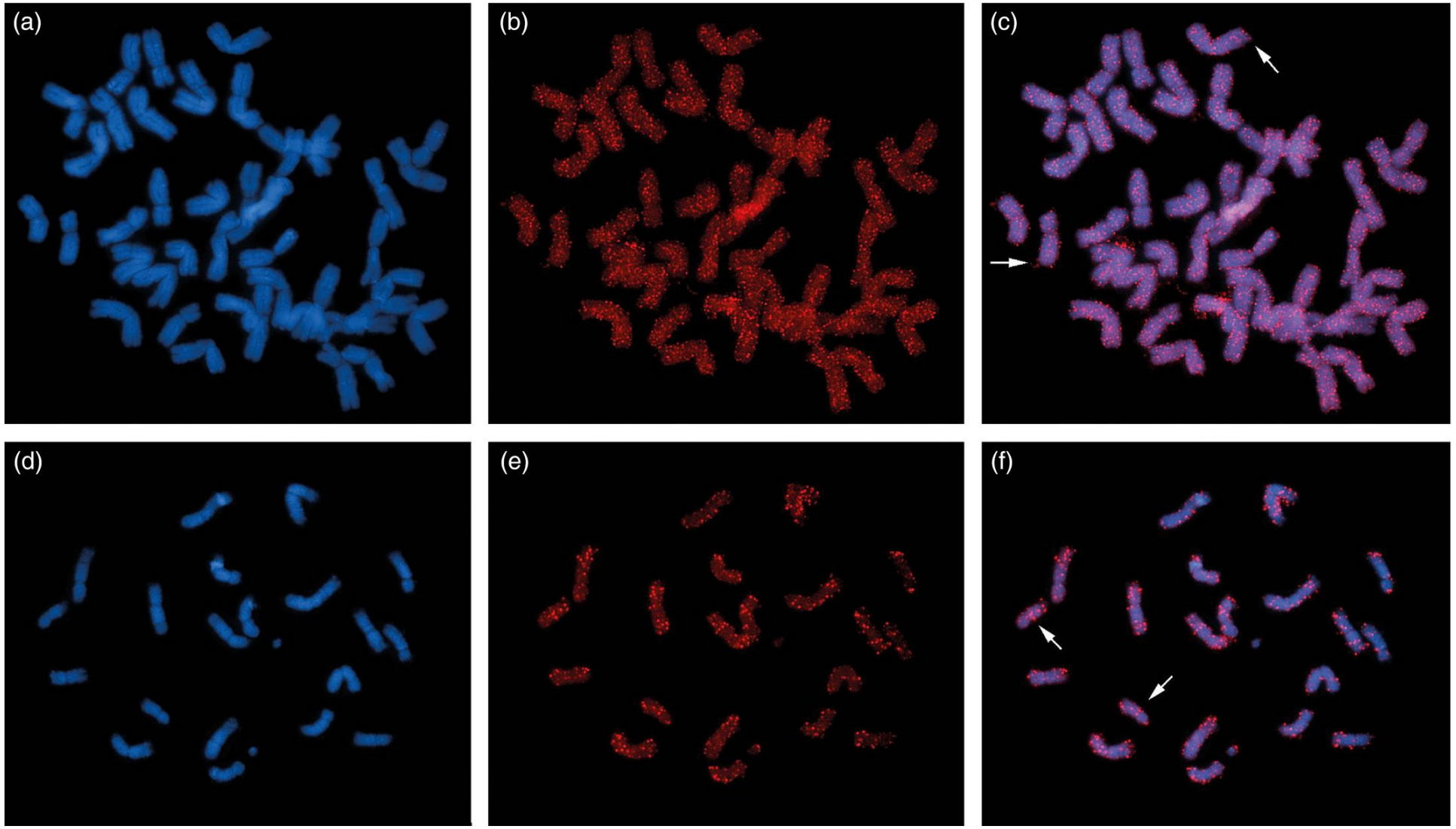


Figure 3. Physical mapping of PoaS-X.2 (*T. aestivum*) and ZmSINE1 (*Z. mays*) SINE copies on metaphase chromosomes. Blue fluorescence (a, d) shows 4′,6-diamidino-2-phenylindole-stained DNA and red signals (b, c, e, f) are sites of SINE hybridization. In *T. aestivum* (b, c) SINEs are uniformly distributed along chromosomes up to the distal regions (arrows). In *Z. mays* (e, f) SINEs are largely clustered in distal and some centromeric regions. Examples of centromeric accumulation are marked by arrows.

Species distribution as indicator of the evolutionary minimum age of grass SINE families

The comparative investigation of the genomic abundance of the 32 SINE families in Poaceae species revealed that the copy number can vary up to three orders of magnitude across species (Figure 1). Moreover, based on the phylogenetic relationships of the grass species investigated and the distribution pattern of the SINE families, we have inferred the minimum age of SINE families (Figure 1). The SINE distribution patterns are patchy across the species investigated (e.g. p-SINE3, OsSN1, OsSN2.2, ZmSINE2.1, ZmSINE2.2, PoaS-III, and PoaS-XII) and do not fully mirror the phylogenetic relationships. As SINEs are propagated unidirectional by the copy-and-paste mechanism, the absence in a certain species is most likely caused by a lack of SINE activity over a long period and subsequent divergence of existing SINE copies until decay (Schwichtenberg *et al.*, 2016; Fawcett and Innan, 2016).

The copy numbers of the SINE families, separated into full-length and 5' truncated elements, are presented in a matrix, which relates the data to the phylogenetic relationship of the seven Poaceae species of the Ehrhartoideae, Pooideae and Panicoideae (Figure 1). The number of SINE families per species ranges from five in maize up to twelve in rice with extensive differences in copy numbers (e.g. p-SINE1 vs. p-SINE3, Table 1). The highest copy number across all species was found for OsSN2.2 (2,685), followed by Au (1,350), and ZmSINE1 (1,111) (Table 1).

The genomes of the closely related Pooideae species wheat and barley separated 13 million years ago (mya) (Gaut, 2002) and largely contain the same SINE families (Figure 1). In contrast, the common ancestor of maize and sorghum millet dates back only 9 mya (Gaut, 2002), but these species show clearly different sets of SINE families, which is presumably the result of lineage-specific evolutionary divergence.

Moreover, although species of the Ehrhartoideae, Pooideae and Panicoideae separated 60 mya (Charles *et al.*, 2009), the conserved SINE families ZmSINE2.1 and ZmSINE2.2 are still present in single species of the three lineages indicating the longevity of some SINE families (Figure 1). Other SINE families are restricted to a single lineage or species of the Poaceae only. For example, ZmSINE3 is distributed in the Panicoideae species *P. virgatum*, *S. bicolor*, and *Z. mays* (Figure 1, red) indicating

lineage-specific amplification approximately 28 mya, while ZmSINE2.3 copies are only found in maize.

The large evolutionary distance between rice and Pooideae species, having the last common ancestor 50 mya (Charles *et al.*, 2009), is reflected by the existence of six SINE families (p-SINE2, OsSN3, PoaS-IV, PoaS-VII, PoaS-IX, and PoaS-XIII) occurring exclusively in rice but not in the Pooideae. Likewise, 16 of the 20 SINE families and subfamilies occurring in the Pooideae do not exist in rice. These data suggest that SINE diversification and amplification proceeded after separation of the Pooideae species from rice. Hence, the group of Pooideae-specific SINE families (PoaS-I, PoaS-II, PoaS-V.2, PoaS-VI, PoaS-X.1-3, PoaS-XI.1-3, and PoaS-XIV) may have arisen between 50 and 35 mya. The SINE families limited to wheat and barley (PoaS-VI, PoaS-X.1-3 and PoaS-XI.1-3) might have emerged even less than 35 mya. PoaS-XIV occurs exclusively in wheat and represents a relatively young and presumably still emerging SINE family.

In contrast, the ancient and widespread Au SINE (Yasui *et al.*, 2001) is present in six of seven analyzed species and probably exists for at least 50 million years in the grasses investigated (Figure 1). The highest copy number was detected in *T. aestivum*, which is closely related to *Aegilops umbellulata*, where Au was first identified.

Regarding genome colonization OsSN2.2 was the most successful SINE family (Figure 1). It is present with 1,250 and 1,155 copies in wheat and barley, respectively (Table S3). In particular, the 5' truncated copies of OsSN2.2 account for the high abundance in the Pooideae species. Taking into account only full-length SINE copies, p-SINE1 has the highest copy number (631) among all Poaceae SINE families. A widespread distribution with moderate copy numbers was observed for the PoaS-V subfamilies PoaS-V.1 and PoaS-V.2, together populating all analyzed Poaceae species.

Similarity intervals indicate periods of transpositional activity

Evolutionarily ancient SINE families have more diverged TSDs and a lower sequence similarity among copies caused by accumulation of mutations over time. However, sudden transpositional bursts must be taken into account as an important amplification mode of SINEs and result in a large number of highly similar copies. Since the consensus sequence reflects the most common primary structure of

all members of a SINE family, nucleotide changes in SINE copies are suitable to evaluate the genetic diversity and the time passed since periods of transpositional activity.

We comprehensively analyzed the sequence similarity of the 32 SINE families across species to monitor periods of SINE activity. We performed a pairwise comparison of SINE full-length copies to the respective family consensus sequences and grouped them into intervals from 60 % to over 90 % similarity. These histograms provide information about the transpositional activity in different Poaceae species (Figure 4, Figure S7).

For example, the PoaS-XIII family has a decreasing number of copies per interval spanning from 60 % to 90 % similarity and hence is considered as continuously active over a long period with a slow decrease of transpositional activity over time in rice (Figure 4a). A recent transpositional burst, recognizable by highly similar copies, for example ranging between 92 % and 100 % similarity, and narrow peaks in the histogram is correlated with a strong amplification, as shown exemplarily for OsSN2.2 in *S. bicolor* (Figure 4a). Multiple transposition periods are proposed for ZmSINE2.1 in *B. distachyon* (Figure 4a), OsSN2.1 in *O. sativa* or PoaS-XII in *T. aestivum* (Figure S7). Further examples of recent transposition are frequently found in wheat, for example ZmSINE1 (136 of 200 copies between 92 % and 100 % similarity) (Figure 4b), Au (328 of 471 copies between 90 % and 100 % similarity), PoaS-X.1 (51 of 53 copies between 92 % and 100 % similarity), and PoaS-X.2 (all copies between 92 % and 100 % similarity) (Figure S7). Consistently, activity profiles of wheat PoaS families correlate clearly with the number of transcribed SINE sequences obtained by *BLAST* searches against the NCBI transcriptome shotgun assembly of *Triticum aestivum* (Figure 4c, Figure S7, Table S8).

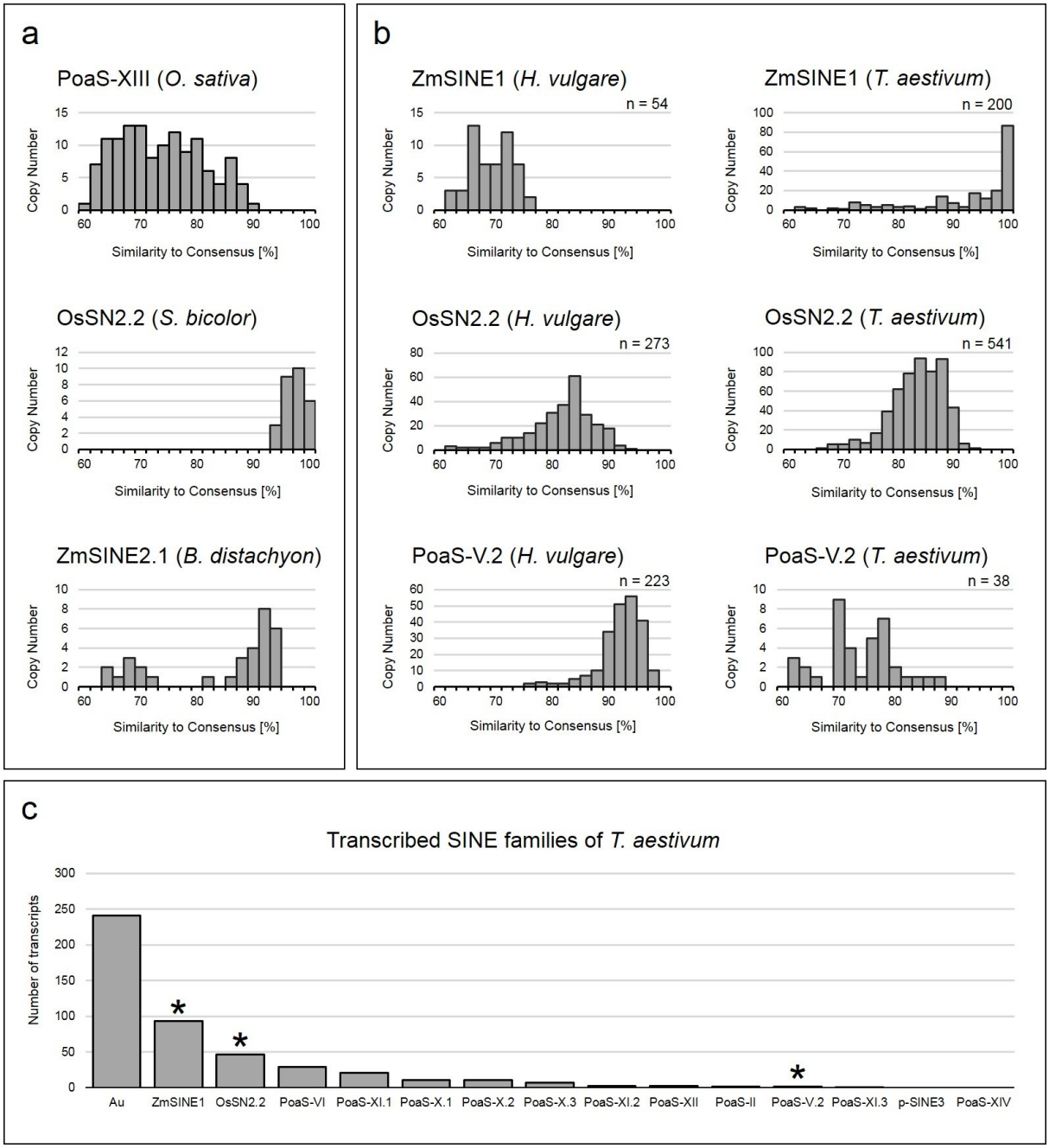


Figure 4. Analysis of transpositional bursts and their influence on abundance based on similarity intervals with assigned copy numbers and transcriptome data. (a) Different patterns of transpositional activity are shown: activity over a long period (PoaS-XIII in *O. sativa*); recent transpositional burst (OsSN2.2 in *S. bicolor*) and multiple transpositional bursts (ZmSINE2.1 in *B. distachyon*). (b) Examples of species-specific amplification of short interspersed nuclear elements (SINEs) in *H. vulgare* and *T. aestivum* shown for ZmSINE1, OsSN2.2 and PoaS-V.2 (number of full-length copies indicated). (c) Number of SINE transcripts (query coverage of at least 80 %) of wheat SINE families and subfamilies. SINEs families shown in (b) are indicated by stars.

The transpositional activity of SINE families has a strong impact on abundance: A fourfold amplification has been observed for ZmSINE1 in *T. aestivum* (136 of 200 copies between 92 and 100 % similarity) compared to *H. vulgare* (54 copies) with lower similarity values (Figure 4b). The same applies to ZmSINE1 copies in *S. bicolor* (23 copies) and *Z. mays* (294 copies) (Figure S7). Despite the increase in copy number in different grass species we did not observe major changes of the ZmSINE1 structure.

OsSN2.2 copies of *H. vulgare* and *T. aestivum* are largely of the same age and the respective copy numbers do not differ dramatically (Figure 4b). In contrast, the PoaS-V.2 copy number in *H. vulgare* (158 of 223 copies between 90 % and 98 % similarity) is almost sixfold higher than in *T. aestivum* (38 copies) (Figure 4b). The burst is accompanied by an 11 bp insertion in the 3' region of PoaS-V.2 in *H. vulgare*, which is missing in *T. aestivum* and *B. distachyon* copies (Figure S8).

Multimerization creates large SINEs

The lengths of all 32 Poaceae SINE families fall into two distinct size ranges (Figure 5). The majority (23) of Poaceae SINE families and subfamilies belong to the length category of 100 bp to 180 bp, whereas the remaining Poaceae SINEs are between 240 bp and 340 bp long. As the size of 240 bp to 340 bp is rather unusual, we examined these SINE families and subfamilies in more detail.

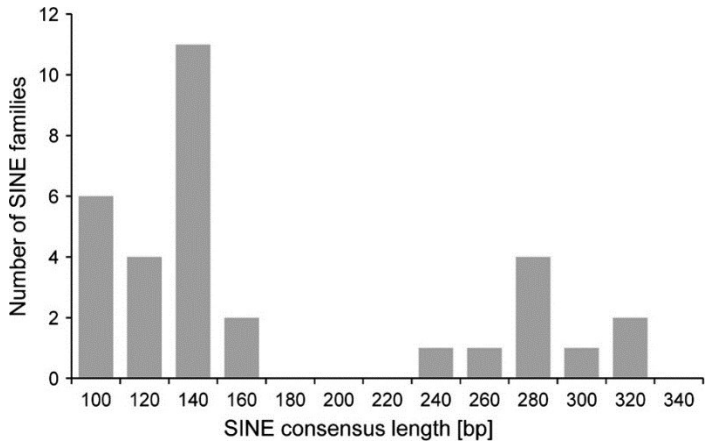


Figure 5. Length distribution of Poaceae SINE families. The length intervals comprise 20 bp (interval starts are indicated). The range is from 108 bp to 333 bp with two maxima of 140 bp - 160 bp and 280 bp - 300 bp, respectively.

We found evidence for the emergence of enlarged heterodimeric SINEs, formed by combination of full-length or nearly full-length SINE copies (Figure 6, Table S9). Most importantly, the combined SINEs are terminated by either poly(A) or poly(T) tails and flanked by TSDs providing evidence that they have indeed been active as multimers. This is consistent with the intact structure of the 5' unit of the multimerized SINEs which is crucial for transcription.

The three ZmSINE2 subfamilies (ZmSINE2.1, ZmSINE2.2, and ZmSINE2.3) as well as PoaS-XIII, PoaS-XIV, PoaS-VII, and the two OsSN2 subfamilies (OsSN2.1 and OsSN2.2) contain internal fusion sites resembling poly(A) tails, poly(T) tails or poly(AC) tails, respectively, which separate the adjacent SINE copies. The RNA polymerase III promotor motifs box A and box B are typically between 31 and 41 nucleotides apart (Figure S4). The sequence of the promotor boxes, their conserved position and distances to each other are significantly more degenerated (designated A' and B') in the 3' SINE units of the dimerized SINE families. In particular, the box A motifs of the 3' SINE units of OsSN1, OsSN2.1, OsSN2.2, and ZmSINE2.1 are strongly diverged and fall below the level of detection (Figure 6).

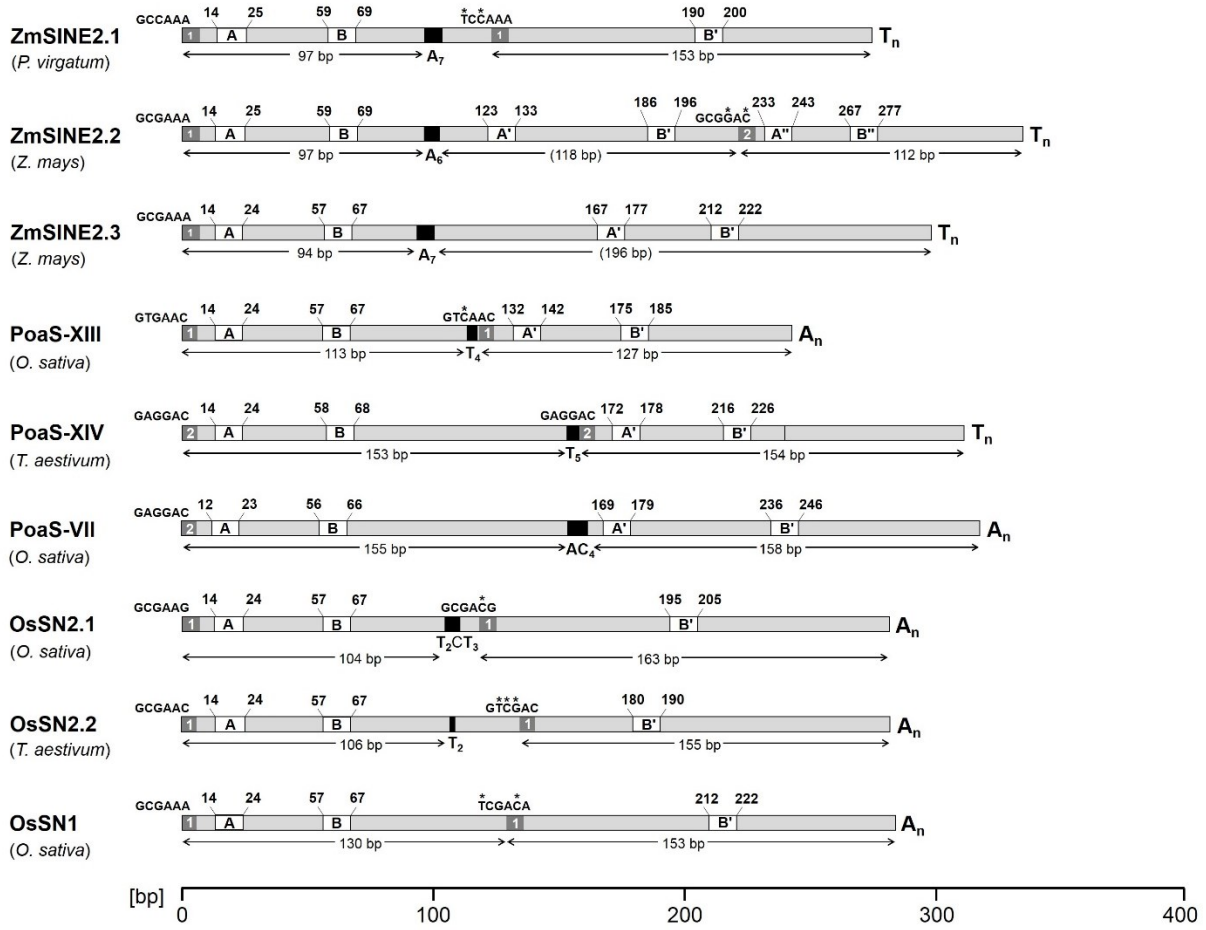


Figure 6. Multimeric Poaceae short interspersed nuclear element (SINE) families. The lengths of the subunits of the SINE multimers were estimated (arrows with size information). The conserved box A and B motifs of the Poaceae SINE families (box A, TAGCNCAG(N)TGG; box B, GGTTCGANNCC; Figure S4) are shown as white boxes with their nucleotide positions within the SINE consensus sequence. Black boxes represent the A-, AC- or T-rich fusion site between the 5′ and 3′ SINE units. Dark grey boxes indicate the first six nucleotides of the 5′ and 3′ SINE units, referring to three typical 5′ start motifs (numbered 1 or 2) of Poaceae SINE families (Figure S5). Deviating nucleotides in the start motif of the 3′ SINE unit, compared with the start motif of the 5′ SINE unit are marked by a star.

Further evidence for SINE multimers is the occurrence of the conserved 5' start sequence motifs, which we have identified as a typical structure for the Poaceae SINEs. Regarding the first six nucleotides of the 5' start, Poaceae SINE families and subfamilies can be assigned to one of the three groups: 5'-GMGAA(M)-3', 5'-GAGGA(M)-3', and 5'-GAAGGG-3' (Figure S6).

In the heterodimer PoaS-XIII, the motif 5'-GMGAA(M)-3' of the 3' SINE unit is located prior the box A' and directly downstream of the fusion site, which consists of four thymines resembling the poly(T) tail of the 5' SINE unit (Figure 6). Furthermore, the spacing between box A' and box B' motif of the 3' SINE unit corresponds to the most common distance of 33 bp (Figure 6, Figure S4). Therefore, PoaS-XIII evolved by integration of a full-length SINE copy downstream of the poly(T) tail of an existing SINE copy. A similar arrangement has been detected in PoaS-VII, although the 3' SINE unit lacks a clear detectable 5' start motif and the distance between box A' and B' is extended.

In OsSN2.1, OsSN2.2, and ZmSINE2.1, the 5' start motif exhibits longer distances to the fusion sites, indicating an integration closely downstream to the 3' tail of a SINE, whereby a short genomic sequence of the 3' flanking region is probably captured in the dimerized SINE. The longest SINE family ZmSINE2.2 (333 bp) constitutes a trimer, as we identified an additional, third promotor motif. In the trimeric ZmSINE2.2 and dimeric ZmSINE2.3, the internal (118 bp) and 3' region (196 bp), respectively, consists of genomic DNA which resembles highly diverged 3' SINE units as we detected the A' and B' box motifs.

Exclusively in wheat, PoaS-XIV constitutes a recently evolved homodimeric SINE (Figure 6), consisting of two tandemly arranged PoaS-X.1 copies, which differ only by two single nucleotide changes (a deletion at position 125 of the 5' unit and a thymine to cytosine transition at position 48 of the 3' unit). The two SINE units of PoaS-XIV (153 bp and 154 bp each) are separated by an internal T-stretch of 5 bp resulting in a consensus length of 312 bp and termination by a 3' tail of 9 thymines. We detected only a single full-length copy which is flanked by a 16 bp TSD, but ten 5' truncated copies and six aberrant fragments (Figure S9).

Evolutionary relations between Poaceae SINE families

The evolution of SINE families is substantially driven by the transpositional activity and diversification. To uncover evolutionary patterns of emergence and divergence, we performed pairwise comparisons of the consensus sequences of all Poaceae SINE families and subfamilies. Only regions with sequence similarities of at least 70 % spanning at least 30 bp were taken into account and considered to be of the same origin (Figure 7, Table S2).

Surprisingly, 28 of 32 Poaceae SINE families and subfamilies are structurally related across Poaceae species and share sequence regions with at least another SINE family or subfamily. Only the SINE families PoaS-I, PoaS-IV, PoaS-VIII and ZmSINE3 did not show any structural relatedness to other grass SINEs in this study.

The lengths of highly similar sequence motifs range between 31 bp (PoaS-VI and PoaS-V.2) and 201 bp (OsSN1 and OsSN2.2), and similarities for corresponding portions were found from 71 % between PoaS-VII and PoaS-XIII up to 99 % between PoaS-XIV and PoaS-X.1 (Figure 7).

Based on the region in which the similarity was found we suggest the following routes of SINE evolution in grasses: Integration of full-length or truncated SINEs from abortive transcripts (‘reshuffling’) (I), diversification and vertical transmission (II), and recombination (III).

(I)

Abortive reverse transcription of full-length SINE copies or reverse transcription of 5’ truncated SINE transcripts, both followed by integration into existing SINEs, might presumably be the most frequent process responsible for partial structural conservation. We postulate that some ancient, highly abundant SINEs such as the OsSN and the ZmSINE families, were target sites for single or multiple integration events of truncated unrelated SINEs thereby resulting in novel chimeric SINEs.

Several groups of SINE families were observed which show considerable similarity in their 5’ or 3’ regions, but variability in the remaining regions:

The three p-SINE families from rice show different 3’ regions and terminate with poly(T) tails. They were most likely formed by the acquisition of different genomic sequences to the same founder SINE

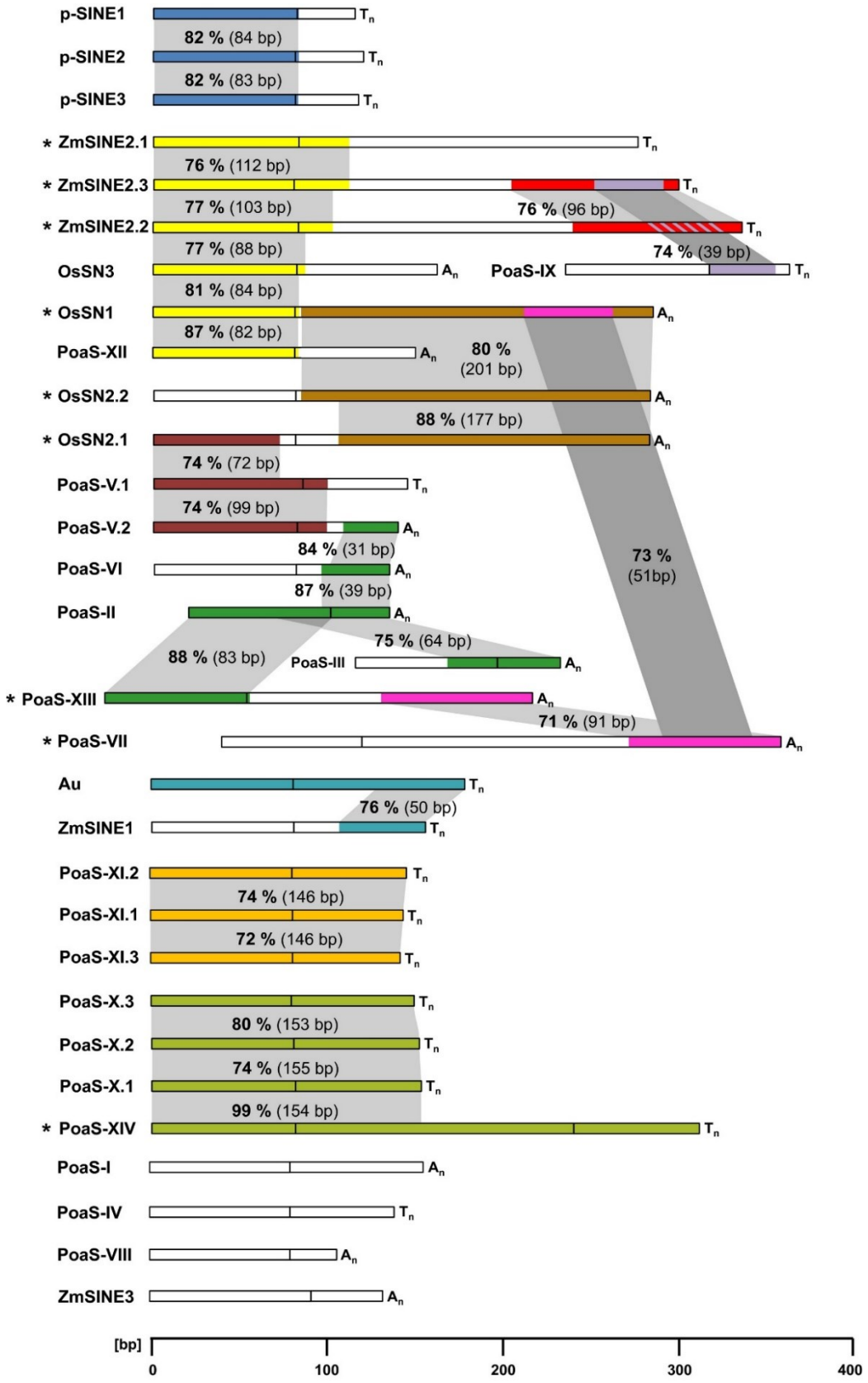


Figure 7. Structural relationships of Poaceae SINE families and subfamilies. Sequence similarities (at least 70 % sequence similarity over a length of at least 30 bp) are shown by identical colors. The 3′ region of ZmSINE2.2 contains an area shown with light purple lines indicating remnants of the 39 bp region present in PoaS-IX and ZmSINE2.2 (similarity below 70 %). Grey shadings illustrate highly similar SINE regions containing the percentage and length of sequence similarity. A black vertical line within the SINE marks the end of the tRNA-related SINE portion. Multimeric SINE families are marked by a star.

family, probably also by integration of truncated members of unrelated SINE families, thus resulting in common 5' regions which show 82 % similarity (Figure 7, blue). Another group is formed by the families ZmSINE2.1, ZmSINE2.2, ZmSINE2.3, OsSN1, OsSN3 and PoaS-XII sharing considerable parts of their 5'regions (82 bp - 112 bp, 76 % - 86 % similarity, Figure 7, yellow), but showing highly variable 3' regions indicating insertion of different truncated SINEs. However, ZmSINE2.3 and ZmSINE2.2 are diverged from ZmSINE2.1 and have a closer relation in their 3' regions, resembling the 3' end of PoaS-IX (Figure 7, purple) and reached a higher complexity.

To demonstrate the relation of the common 5' regions, a rooted dendrogram was constructed based on representative sequences of the conserved motifs that these SINEs have in common (Figure 8). Copies showing the highest similarity to the consensus element were selected and only the first 80 bp of their 5' end were analyzed. The 5' ends of ZmSINE2.3 form a distinct clade, while the ZmSINE2.1 and ZmSINE2.2 sequences are grouped together indicating a more recent emergence of these subfamilies, as their 5' regions gained less characteristic mutations yet (Figure 8). In contrast, the 5' ends of ZmSINE2.3, OsSN3, OsSN1, and PoaS-XII form family-specific clades.

OsSN1 of this group is linked with OsSN2.2 and OsSN2.1 by sharing a large part of the 3' region indicating that these OsSN families probably emerged by the integration of copies from the same SINE family (Figure 7, brown). Moreover, OsSN1 is a composite SINE which carries additionally 51 bp of the 3' part of PoaS-VII (Figure 7, pink). The 3' SINE unit of the heterodimeric PoaS-VII family is also found in PoaS-XIII: Both share 91 bp of their 3' end including the poly(A) tail.

A group of SINEs is related with PoaS-II: It shares the 3' region with PoaS-VI and PoaS-V.2 and leading in the latter to the donation of a poly(T) tail. PoaS-II shares also a large part of its sequence (75 % similarity over 64 bp) with PoaS-III. However, it remains unclear which SINE family was the founder of this group, since PoaS-II, PoaS-III and PoaS-VI can be taken into account as donor SINE. The heterodimer PoaS-XIII is also a chimeric SINE as it contains the same 5' SINE region as PoaS-II.

The similarity of 76 % over 50 bp of the widespread Au SINE with ZmSINE1 could be explained by integration of a 5' truncated Au copy into the 3' region of a precursor SINE of ZmSINE1 (Figure 7, turquoise).

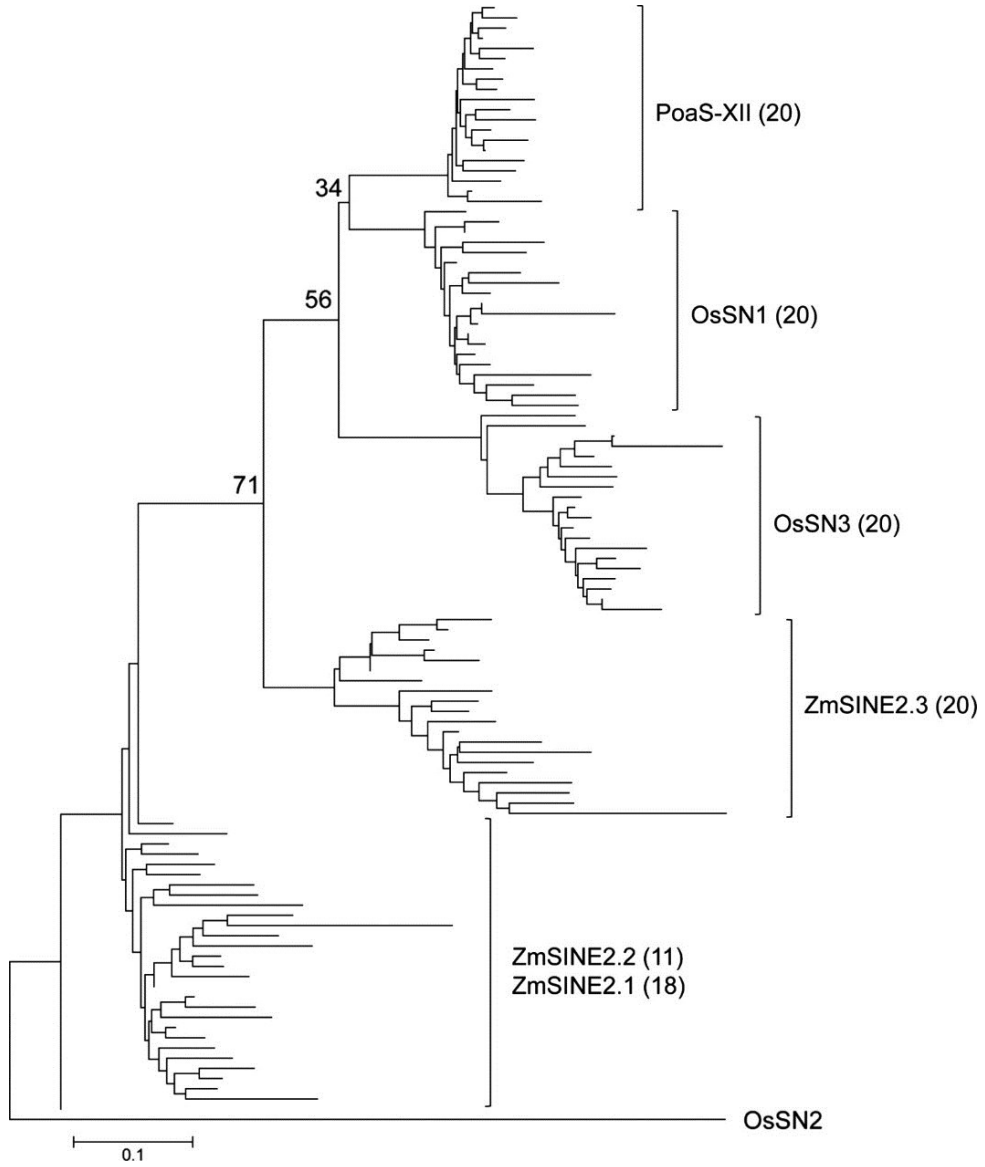


Figure 8. Phylogenetic relationship of a subclade of SINEs sharing the same 5′ region. The first 80 bp of the 5′ end of ZmSINE2, OsSN1, OsSN3 and PoaS-XII are highly similar. Up to 20 representative 5′ end sequences (80 bp) of each SINE family, exhibiting the highest similarity to the consensus element, were used for the construction of the dendrogram. The OsSN2 consensus element was used as an outgroup sequence. The nucleotide divergence scale is indicated below.

(II)

PoaS-XI.1, PoaS-XI.2, PoaS-XI.3, PoaS-X.1, PoaS-X.2, and PoaS-X.3 diversified vertically by accumulation of single nucleotide mutations or small indels without large structural changes resulting in subfamily structures. These SINEs occur only in the evolutionarily closely related grasses barley and wheat, and only differ in short regions with a maximum length of five nucleotides or by single diagnostic nucleotide exchanges, respectively (Figure 7, orange and pale green).

(III)

Other structural relationships among PoaS families are based on similar internal regions, observed close to the 3' ends of SINEs (PoaS-VII and OsSN1, ZmSINE2.3 and PoaS-IX) (Figure 7, pink and purple). ZmSINE2.3 shares a 39 bp region with PoaS-IX 8 bp prior to the 3' tail. After acquisition of the 3' region of PoaS-IX, the outermost 3' end (8 bp) of ZmSINE2.3 might have been replaced by an extremely short ZmSINE2.2 region (8 bp and the 3' tail). Alternatively, it might have been diverged over time or the result of recombination (e.g. template switch). ZmSINE2.2 also includes a PoaS-IX portion, however, similarity is below 70 % (purple dotted lines in Figure 7).

Species-specific diversification forms subfamilies

OsSN2.1 might have possibly been the donor of the 5' region of the subfamilies PoaS-V.1 and PoaS-V.2 differing in their 3' regions (Figure 7). In PoaS-V.2, the 3' poly(A) tail can be traced back to a truncated copy of PoaS-II, PoaS-III or PoaS-VI. This structural peculiarity is strongly correlated with their contrasting distribution pattern among Poaceae species: Members of PoaS-V.1 are only present in species of the Panicoideae and Ehrhartoideae, while PoaS-V.2 is restricted to the Pooideae including *B. distachyon*, *T. aestivum* and *H. vulgare* (Figure1).

To visualize the interspecific divergence of PoaS-V on the sequence level, an unrooted dendrogram was constructed, using at most 20 representative copies of PoaS-V.1 and PoaS-V.2 of each plant species (Figure S8). Both SINE subfamilies form to two main branches containing PoaS-V.2 SINEs from *B. distachyon*, barley and wheat, and PoaS-V.1 SINEs from rice, switchgrass, sorghum millet and maize (Figure 9).

Moreover, within the main branches, species-specific diversification was observed for both PoaS-V subfamilies: PoaS-V.2 copies of *H. vulgare* are distally positioned on a separate branch, while PoaS-V.2 copies of *B. distachyon* and wheat show only minor differences to each other (Figure S8) and are therefore grouped together.

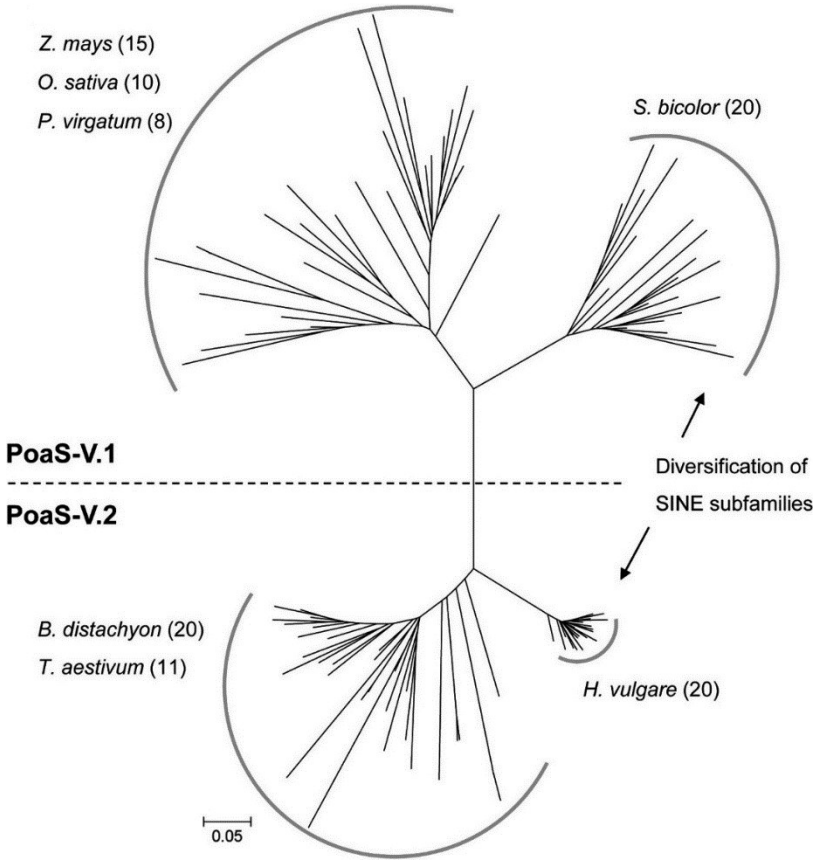


Figure 9. Dendrogram showing the species-specific diversification of the SINE subfamilies PoaS-V.1 and PoaS-V.2. Up to 20 PoaS-V copies of each species with at least 70 % identity to the species-specific SINE family consensus sequence (Figure S8) were used for the construction of the dendrogram.

The short branches within the clade of barley PoaS-V.2 SINEs indicate a high similarity and species-specificity of PoaS-V.2 copies in barley, also accompanied by an 11 bp deletion in the 3' region which is not found in PoaS-V.2 copies of *B. distachyon* and wheat (Figure S8). Remarkably, 192 of 223 PoaS-V.2 copies in *H. vulgare* (Figure 4b) show at least 90 % similarity to the consensus element (Figure 4b) suggesting recent diversification followed by amplification.

Similarily, PoaS-V.1 copies of *S. bicolor* also form a separate clade next to the PoaS-V.1 copies of rice, switchgrass and maize. The *S. bicolor*-specific PoaS-V.1 consensus sequence differs by 40, 25, and 30

diagnostic single nucleotide polymorphisms (SNPs) from those of rice, switchgrass and maize, respectively (Figure S8). PoaS-V.1 copies of switchgrass and maize are more closely related as their consensi show only 14 diagnostic nucleotide changes.

Discussion

Insights into the SINEs of grass genomes

The *de novo* assembly and annotation of large genomes including those of major crops still remain a challenging and laborious task caused by the large repetitive fraction of plant genomes (International Rice Genome Sequencing Project, 2005; Schnable *et al.*, 2009; Arabidopsis Genome Initiative, 2000; Vogel *et al.*, 2010; Mayer *et al.*, 2012; Brenchley *et al.*, 2012). Therefore, the identification and characterization of repetitive sequences is crucial for genome annotation, while conversely genome sequences enable the understanding of repeat organization and evolution. In particular, genome sequences are an excellent resource to gain knowledge of small but abundant retrotransposons such as SINEs which have only been comprehensively investigated in rare cases in plants (Lenoir et al., 1997; Lenoir *et al.*, 2001; Lenoir *et al.*, 2005; Deragon and Zhang, 2006; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016; Seibt *et al.*, 2016).

However, the annotation of SINEs gains significance (Dohm *et al.*, 2013; Aversano *et al.*, 2015; Vu *et al.*, 2015; Jiao *et al.*, 2017). In this study, we retrieved 11,052 SINEs falling into 32 Poaceae families and subfamilies, which is the highest number of SINE families characterized in a plant family so far. In the Amaranthaceae 22 SINE families have been recently described, while in the Brassicaceae 16 SINE families and in the Fabaceae 15 SINE families are known (Lenoir *et al.*, 1997; Deragon and Zhang, 2006; Gadzalski and Sakowicz, 2011; Schwichtenberg *et al.*, 2016). In the Solanaceae ten SINE families and subfamilies with more than 82,000 copies have been characterized (Wenke *et al.*, 2011; Seibt *et al.*, 2016).

Highly differing copy numbers of SINE families in the Poaceae species investigated are the result of the copy-and-paste amplification by retrotransposition. Hence, the SINE populations observed are a snapshot of the situation between periods of amplification and gradual degeneration by mutations (Schwichtenberg *et al.*, 2016; Fawcett and Innan, 2016). TSDs flanking the SINEs are also subject to mutations, and thus, their unambiguous determination and delimitation is often difficult, and in many studies the TSDs were excluded from detailed analysis (Lenoir *et al.*, 2001; Deragon and Zhang, 2006; Tsuchimoto *et al.*, 2008; Baucom *et al.*, 2009; Wenke *et al.*, 2011). We observed a statistically significant positive correlation of average TSD lengths with average similarities of the full-length

SINE sequences (Figure 2, Figure S2) implying that they might be suitable as indicators for recent activity and insertion. The TSD lengths of Poaceae SINE families (average 6 bp - 16 bp, 24 bp maximum) are variable (Figure S1, Table S6), and in a similar range as reported for the TSD length of Fabaceae SINEs and Amaranthaceae SINEs which are 9 bp - 20 bp and 7 bp - 13 bp in size, respectively (Gadzalski and Sakowicz, 2011; Schwichtenberg *et al.*, 2016). However, some SINE copies in Amaranthaceae species have extreme TSDs reaching up to 36 bp (Schwichtenberg *et al.*, 2016).

The majority of plant SINEs terminate with a poly(A) tail. Surprisingly, among the 32 SINE families, we identified 18 families and subfamilies with a poly(T) tail, which presumably might be a specific feature of SINEs in grasses. Moreover, all poly(T) SINEs described so far in plants are restricted to and specific for the Poaceae suggesting that this motif emerged at least 60 mya in the last common progenitor and has presumably contributed to the successful propagation of the respective SINE families (Umeda *et al.*, 1991; Yasui *et al.*, 2001; Xu *et al.*, 2005; Baucom *et al.*, 2009). An exception is the Au SINE, first detected in the wheat progenitor *Aegilops umbellulata*, that also terminates by a poly(T) tail, but is not restricted to grasses and widespread in angiosperms and gymnosperms suggesting its emergence 200 mya (Yasui *et al.*, 2001; Fawcett and Innan, 2016). However, most genomes, in particular rice and wheat genomes contain also poly(A) terminating SINE families. The average and maximum length of the poly(A) or poly(T) tails are similar to that of most plant SINEs, but do not reach the extremes observed in Solanaceae SINEs and Amaranthaceae SINEs which are characterized by tails with up to 45 and 48 adenines, respectively (Wenke *et al.*, 2011; Seibt *et al.*, 2016; Schwichtenberg *et al.*, 2016). In animals, extended tails covering more than 40 residues, have been associated with recent activity and insertion (Odom *et al.*, 2004). Although some Poaceae SINE families must have been active in the recent past due to a large number of highly similar copies, extended 3' tails were not found for these SINEs. However, it remains unclear if the extended 3' tail of a SINE copy is transcribed and also integrated completely by the LINE-RT or if it is just required for higher stability upon binding to the target site during target-primed reverse transcription or for recognition and binding of the relevant proteins, so that parts of the tail can get lost. Also, the observed negative correlation (correlation factor of - 0.18 and p-value of 0.32) between the average tail lengths

and the similarity of Poaceae SINE families (Figure 2, Figure S2) indicates no clear trend for a successive shortening over time as suspected for TSD lengths and similarity. However, despite the fact that SINE tails are also subject to an ongoing accumulation of mutations, their original length can be estimated using the positions of the conserved 3' ends and the TSDs.

Plant SINEs, including the SINE families identified in this study, are derived from tRNA genes. The tRNA-derived region is relatively conserved in size and terminates shortly (14 nucleotides) after the box B motif (Deragon and Zhang, 2006). Hence, the length variation of SINE families is mainly determined by the 3' region. Although SINEs up to 500 bp in length have been described (Kajikawa and Okada, 2002), most plant SINEs are typically 100 bp to 250 bp long. In Poaceae species, 72 % of all SINE families and subfamilies belong to the size category of 100 bp to 180 bp with PoaS-VIII being the smallest (108 bp). Similarily, Fabaceae SINEs are between 140 bp and 200 bp in length, Solanaceae SINEs range between 106 bp and 244 bp, and Amaranthaceae SINEs between 113 bp and 223 bp (Gadzalski and Sakowicz, 2011; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). Longer Poaceae SINEs, exemplified by ZmSINE2.2 (333 bp), are the result of structural rearrangements and fall in a second size category of 240 bp to 340 bp. A broader size range was found in Brassicaceae SINEs with 95 bp (SB8) up to 352 bp (SB7) (Deragon and Zhang, 2006). The shortest SINE family described so far was detected in *Manihot esculenta* with 83 bp (EuphS-I) (Wenke *et al.*, 2011). This suggests that only the tRNA-related portion together with the 3' tail is required to form a minimalistic but functional and transposition-competent SINE (e.g. DAS-Ia, Churakov *et al.*, 2005). Hence, any genomic sequence may become part of a SINE, provided that it is situated between the promotor motif and an adenine or thymine stretch, fulfilling the function of a 3' tail, not more than approximately 400 bp downstream and containing a sequence region with similarity to the transcriptional terminator motif (Comeaux *et al.*, 2009). This length constraint is in line with the elongation rate of RNA polymerase III (Schramm and Hernandez, 2002).

Species-specific transpositional activity results in highly diverse SINE landscapes

For the existence over long evolutionary time scales, at least a single copy of a SINE family ('master copy') has to have a 'safe' genomic environment ensuring its intactness and transposition competence

(Deininger *et al.*, 1992; Schwichtenberg *et al.*, 2016; Fawcett and Innan, 2016). The accumulation of mutations depends on the time passed after transposition, in addition to various factors such as genomic and chromosomal position. Therefore, SINE families with recent or ongoing transposition harbor more homogeneous copies.

We found that the transpositional activity of SINE families is variable in scale and duration and independent of the Poaceae species, resulting in species-specific differentiation during the radiation of the Poaceae (Figure 4, Figure S7). For example, ZmSINE1 lost its activity in some species (in barley earlier than in maize), while it is presumably still active in wheat suggested by the high number of homogeneous copies (Figure 4, Figure S7). Similar observations are reported from SINEs in Amaranthaceae and, in particular for potato and tomato in the Solanaceae (Schwichtenberg *et al.*, 2016; Seibt *et al.*, 2016). However, similar activity profiles of SINEs across species borders were detected between cultivated and wild varieties of tomato (Seibt *et al.*, 2016).

Moreover, we also found indication for the reactivation of a SINE family after a long period of inactivity: OsSN2.2 must have emerged in an ancestor of the Poaceae 60 mya, since it is distributed in species of the Pooideae and Panicoideae (Figure 1). While the transpositional activity of OsSN2.2 has ceased in Pooideae species such as barley and wheat, homogenous and hence relatively recently inserted copies were found only in *S. bicolor* (Figure S7).

The activity profiles deduced from similarity intervals allow insights into the origin of subfamilies, e.g. in the group of PoaS-X subfamilies: PoaS-X.3 has been active for a long period in barley and wheat, while PoaS-X.1 and PoaS-X.2 emerged later and were amplified to a different extent in both species (Figure S7). Thus, PoaS-X.1 and PoaS-X.2 most likely evolved from diversified PoaS-X.3 copies.

It has to be taken into account that the number of SINE copies may be underestimated due to too high diversification (similarity to consensus falling below 60 %). Moreover, our conclusions about activity profiles rely on similarity values which are based on the assumption of an equal mutation rate over time and in different genomic regions or dependent on the chromatin status. Nevertheless, the majority of wheat SINE families may still be active as a high number of transcripts was detected in transcriptome data (Figure 4c, Table S8). Presumably, not all transcripts must originate from SINE

activity, since SINEs are frequently found in genes or genic regions (Lenoir *et al.*, 2001; Baucom *et al.*, 2009; Seibt *et al.*, 2016) and, therefore, are not necessarily transcribed by RNA polymerase III. In wheat, the highest number of SINE transcripts (241, Table S8) was found for the ancient and widespread Au family which indicates, together with a high number of young copies (328 of 471 copies between 90 % and 100 % similarity, Figure S7), that it might represent the currently most successful propagating SINE in the genome of *T. aestivum*. Furthermore, this example demonstrates that the relative age of SINE copies gives insights into the recent transpositional behavior, but cannot be correlated with the estimated minimum age of the SINE family.

The physical mapping of SINEs by FISH (Figure 3) revealed a contrasting hybridization along chromosomes with a preferred distal clustering of ZmSINE1 in maize and more uniformly scattered PoaS-X copies along wheat chromosomes. The dispersed distribution of PoaS-X corresponds with the general weak SINE insertion preference, which is specified by only a single adenine or thymine or short stretches thereof (Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). Especially PoaS-X.1 and PoaS-X.2 contain many highly similar copies (96 % - 97 % average similarity, Table 1) suggesting a more recent integration. In contrast, ZmSINE1 consists of more evolutionarily older copies, reflected by 75 % average similarity (Table 1). Hence, the observed accumulation of ZmSINE1 in distal and pericentromeric chromosome regions might be the consequence of a fast SINE turn-over following insertion, opposed to regions providing a safe environment for the survival of SINEs like distal, gene-rich chromosome regions (Seibt *et al.*, 2016; Mascher *et al.*, 2017).

Reshuffling-based evolution as a main route of Poaceae SINE family emergence

We provide evidence that new Poaceae SINE families mainly evolve from existing SINEs detectable by conservation and similarity of 5’ or 3’ parts.

Indications for evolutionarily young SINE families are high similarities between the 5’ region and a specific tRNA gene over the whole length, if the SINE developed *de novo* (Zhang and Wessler, 2005; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). The ’reshuffled’ structure of Poaceae SINE families (Figure 7) suggests different evolutionary scenarios that can explain the conservation and relatedness

between SINE families and their routes of diversification. We postulate the following model of SINE evolution:

SINE emergence based on the abortive reverse transcription of SINE copies or reverse transcription of 5' truncated SINE transcripts (e.g. read-through transcription starting adjacent to genes) into existing SINEs is illustrated in the model shown in Figure 10.

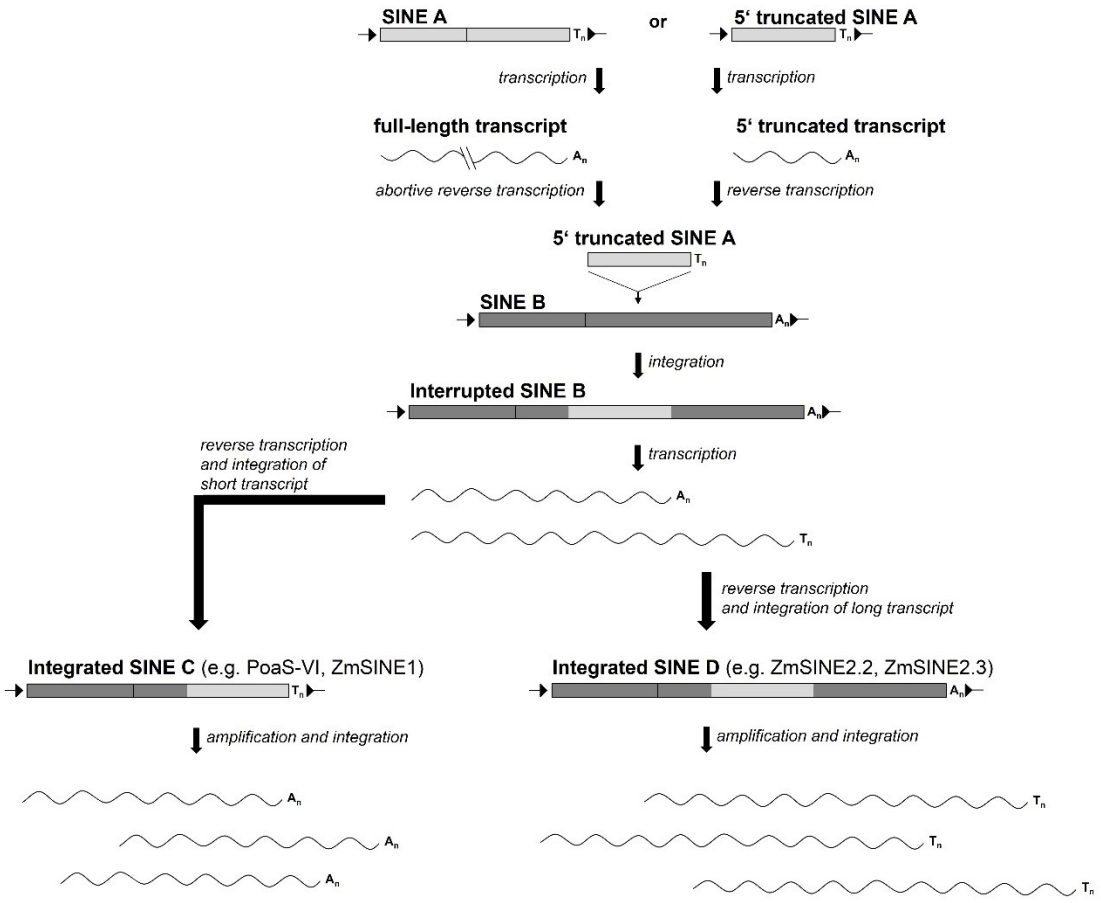


Figure 10. Model for the evolution of reshuffled Poaceae SINEs. A 5′-truncated SINE copy (abortive reverse transcription of SINE A or reverse transcription of a 5′-truncated SINE A) integrates into the 3′ region of SINE B. Activation of the interrupted SINE B copy results in two differing scenarios: reverse transcription starts either at the 3′ tail of the integrated SINE A (internal) or SINE B (3′ end of interrupted SINE B), leading to the new SINEs C and D, respectively. Black triangles indicate target site duplications. A black vertical line within the SINE marks the end of the tRNA-related SINE portion.

This scenario is supported by a high variability observed in the 3' structure of Poaceae SINE families. For example, the different length of the common 5' region of the ZmSINE2 families, OsSN3, OsSN1, and PoaS-XII can be explained by recruitments of diverse sequences in their 3' region (Figure 7, Figure 8). Also, there are SINE families (e.g. ZmSINE2.3, OsSN1) with rearranged 3' regions originating from several different SINE families suggesting nested integration. Consistently, the number of 5' truncated SINE copies exceeds the number of full-length SINE copies in nine SINE families (Figure S3, Table 1; Wenke *et al.*, 2011). Most likely, the truncation is introduced during target-primed reverse transcription which starts at the 3' tail and continues towards the 5' tRNA-related region (Zingler *et al.*, 2005). Interruption of this process occurs frequently and has been described also for LINEs, which provide the reverse transcription machinery for SINEs (Chen *et al.*, 2007; Wenke *et al.*, 2009; Wenke *et al.*, 2011).

Also, 5' truncated LINE transcripts may contribute to the reorganization of the 3' region of SINEs. In a similar way, the formation of the TS SINE in tobacco was explained by the integration of a 5' truncated LINE sequence (SolRTE-1) into the SINE SolS-V (Wenke *et al.*, 2011). Moreover, ZmSINE2 and ZmSINE3 share their 3' end with LINE1-1 (Baucom *et al.*, 2009). Additional SINE/LINE partnerships within the Poaceae were not detected yet.

The phenomenon of reshuffled SINE structures was also reported in the Brassicaceae (Lenoir *et al.*, 1997; Zhang and Wessler, 2005; Deragon and Zhang, 2006), in rice (Tsuchimoto *et al.*, 2008), and in animals (Ziętkiewicz and Labuda, 1996; Buzdin *et al.*, 2002; Buzdin *et al.*, 2003; Takahashi and Okada, 2002; Nishihara *et al.*, 2006). However, chimeric structures were predominantly described for SINE subfamilies rather than for different families in animals and plants (Roy *et al.*, 2000; Takahashi and Okada, 2002; Zhang and Wessler, 2005).

An important scenario is the evolution of large SINEs by adjacent integration of related or unrelated SINE copies resulting in homodimerization (PoaS-XIV) or heterodimerization (remaining examples in Figure 6 with ZmSINE2.2 as a potential trimeric SINE). A striking example for an ongoing emergence of novel SINE families is the single homodimeric PoaS-XIV copy which consists of two former PoaS-X.1 copies: The PoaS-XIV SINE exclusively exists in wheat, while the founder PoaS-X.1 subfamily is present in moderate copy number in wheat and barley.

SINE trimers were only recorded in the colugo (CYN-III; Schmitz and Zischler, 2003) and in the tree shrew (Tu type II; Nishihara *et al.*, 2002) so far. In animals, a broader range of different SINE dimers is known: Both units derived from tRNA (Feschotte *et al.*, 2001; Churakov *et al.*, 2005), hybrid 7SL RNA/tRNA SINEs (Nishihara *et al.*, 2002), and both units derived from 7 SL RNA (Ullu and Tschudi, 1984).

The generation of species-specific SINE variants with diagnostic nucleotide exchanges together with a stepwise and random recruitment of alternative 3' regions results in novel SINE families. All structural rearrangements are followed by amplification and population of the respective genomes. After integration of SINE 3' regions (Figure 10, SINE A) into existing, unrelated SINEs, the 3' tail of the originally intact SINE copy (Figure 10, SINE B) is not needed for reverse transcription of the newly formed SINE and most likely decayed as it is no longer detectable (Figure 10, SINE C). However, due to the presence of two SINE tails in the interrupted SINE B copy (internal T stretch and 3' poly(A) tail), alternative transcripts are possibly contributing to an ongoing diversification (Figure 10, SINE C and D).

Divergence during transmission from generation to generation combined with episodes of amplification over evolutionary time scales results in SINE subfamily structures with numerous diagnostic mutations as it was observed for the PoaS-XI, PoaS-X and OsSN2 subfamilies. This mode of subfamily formation is widespread in plants and animals (Deininger and Batzer, 1995; Price *et al.*, 2004; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016). The changes may either result from spontaneous mutations in the integrated SINE-DNA or introduced during reverse transcription. The reverse transcriptase is error-prone and lacks the proofreading function found in conventional DNA polymerases resulting in error rates which are orders of magnitude higher.

It is also conceivable that gene conversion and template switch are involved in the formation of new SINE families explaining shared internal regions. For example, ZmSINE2.3 shares a 39 bp internal region with PoaS-IX, 7 bp prior to the 3' end (Figure 7). Gene conversion refers to a recombination event between two different SINE copies of the genome. Based on highly similar regions, a SINE copy donates a part of its sequence to another SINE copy, thereby substituting a part of the sequence of the acceptor SINE copy (Ziętkiewicz and Labuda, 1996; Lenoir *et al.*, 1997). In contrast, the

template switch of the reverse transcriptase, described as a common phenomenon for retroviruses and retroelements, is most probably based on RNA recombination during the reverse transcription of multiple cellular RNAs into cDNA (Negroni and Buc, 2001; Bibiłło and Eickbush, 2002; Bibiłło and Eickbush, 2004). Indeed, it could be shown that recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo* based on a multiple template jumping of the LINE-RT (Yadav *et al.*, 2012). The newly generated chimeric SINE has to be propagated to create a novel SINE family. Hence, the template switch model may also explain reshuffled SINE structures, as it is based on the shared retrotranspositional machinery of SINEs and LINEs, involving SINE transcripts, as well as transcribed LINEs and pseudogenes as putative 'switch partners' (Buzdin *et al.*, 2002; Buzdin *et al.*, 2003).

SINE distribution patterns in grass species indicate frequent lineage-specific extinction of families

Plant SINE families are usually distributed within closely related species but can also exhibit surprisingly high levels of partial or complete conservation and similarity over wide taxonomic distances. Among the 32 SINE families and subfamilies, only the PoaS-V SINE family, consisting of the subfamilies PoaS-V.1 and PoaS-V.2 is present in all seven grasses investigated here. Widespread are also ZmSINE1 and Au which were found in six species while eight SINE families (e.g. PoaS-I, PoaS-IV, PoaS-VII, PoaS-IX, and PoaS-XIV) are present only in a single genome.

As SINEs are mobilized as copy-and-paste retrotransposons, copies are retained in the genome. Consistently, conservation of SINE families in very distantly related species as described for the Au SINE family, widely distributed among angiosperms and gymnosperms, can be explained by vertical transmission (Fawcett and Innan, 2016). Similarily, it is conceivable that many SINE families described here, such as SINEs occurring in single species (e.g. PoaS-IX), might have emerged in the ancestor of the Poaceae and before the split of the three Poaceae subfamilies (Pooideae, Panicoideae, and Ehrhartoideae) 60 mya and have been vertically transmitted.

Nevertheless, lineage-specific SINE copy numbers and the 'patchy' distribution of the PoaS families, i. e. inconsistent with the phylogenetic relation of the species analyzed, indicate evolutionary dynamics involving species-specific diversification and amplification.

Although the removal of SINE copies probably occurs in some rare cases and is possibly caused by short genomic deletions or recombination between small homologous regions such as the TSDs (Devos *et al.*, 2002; Van De Lagemaat *et al.*, 2005), the complete loss of all copies of a SINE family in a species while it is conserved in others is very unlikely. Horizontal transfer, perhaps mediated by animal pests or by close physical contact is frequent and an import mode of genome evolution (Bock, 2010; Gilbert *et al.*, 2010; Schaack *et al.*, 2010). However, the wide geographic distribution of the Poaceae species makes this event very unlikely to explain the patchy SINE distribution.

Therefore, it is more conceivable that SINEs became inactive and highly degenerated in some lineages until they fall below the level of recognizability by our approach and escape detection. An alternative scenario underlying the patchy distribution might be incomplete lineage sorting of the active SINE copy during speciation or dramatic structural rearrangements of this 'master copy' giving rise to new chimeric SINE families by reshuffling and thereby replacing former variants.

References

Aliscioni, S., Bell, H.L., Besnard, G., et al. (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C 4 origins. *New Phytol.*, **193**, 304–312.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Aversano, R., Contaldi, F., Ercolano, M.R., et al. (2015) The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell*, **27**, 954–968.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., SanMiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*., **5**, e1000732.

Bennetzen, J.L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.*, **65**, 505–530.

Bibiłło, A. and Eickbush, T.H. (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J. Biol. Chem.*, **279**, 14945–14953.

Bibiłło, A. and Eickbush, T.H. (2002) The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.*, **316**, 459–473.

Bock, R. (2010) The give-and-take of DNA: horizontal gene transfer in plants. *Trends Plant Sci.*, **15**, 11–22.

Boeke, J. (1997) LINEs and Alus - the poly A connection. *Nat. Genet.*, **16**, 6–7.

Brenchley, R., Spannagl, M., Pfeifer, M., et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T. and Sverdlov, E. (2003) The human genome contains many types of chimeric retrogenes generated through *in vivo* RNA recombination. *Nucleic Acids Res.*, **31**, 4385–4390.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y. and Sverdlov, E. (2002) A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of L1. *Genomics*, **80**, 402–406.

Chen, J.M., Férec, C. and Cooper, D.N. (2007) Mechanism of Alu integration into the human genome. *Genomic Med.*, **1**, 9–17.

Churakov, G., Smit, A.F.A., Brosius, J. and Schmitz, J. (2005) A novel abundant family of retroposed elements (DAS-SINEs) in the nine-banded armadillo (*Dasypus novemcinctus*). *Mol. Biol. Evol.*, **22**, 886–893.

Comeaux, M.S., Roy-Engel, A.M., Hedges, D.J. and Deininger, P.L. (2009) Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? *Genome Res.*, **19**, 545–555.

Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, **10**, 691–703.

Deininger, P., Lander, E., Linton, L., et al. (2011) Alu elements: know the SINEs. *Genome Biol.*, **12**, 236.

Deininger, P.L. and Batzer, M.A. (1995) SINE master genes and population biology. In R. Maraia, ed. *The Impact of Short, Interspersed Elements (SINEs) on the Host Genome.* Georgetown, Texas: Landes, R G, pp. 43–60.

Deininger, P.L., Batzer, M.A., Hutchison, C.A. and Edgell, M.H. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.*, **8**, 307–311.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.*, **12**, 1075–1079.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.*, **35**, 41–48.

Dohm, J.C., Minoche, A.E., Holtgräwe, D., et al. (2013) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, **505**, 546–549.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Fawcett, J.A. and Innan, H. (2016) High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. *Mol. Biol. Evol.*, **33**, 2593–2604.

Fawcett, J.A., Kawahara, T., Watanabe, H. and Yasui, Y. (2006) A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant Mol. Biol.*, **61**, 505–514.

Feschotte, C., Fourrier, N., Desmons, I., Mouches, C. and Mouchès, C. (2001) Birth of a retroposon: the twin SINE family from the vector mosquito *Culex pipiens* may have originated from a dimeric tRNA precursor. *Mol. Biol. Evol.*, **18**, 74–84.

Gadzalski, M. and Sakowicz, T. (2011) Novel SINEs families in *Medicago truncatula* and *Lotus japonicus*: bioinformatic analysis. *Gene*, **480**, 21–27.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Gaut, B.S. (2002) Evolutionary dynamics of grass genomes. *New Phytol.*, **154**, 15–28.

Gilbert, C., Schaack, S., Pace II, J.K., Brindley, P.J. and Feschotte, C. (2010) A role for host–parasite interactions in the horizontal transfer of transposons across phyla. *Nature*, **464**, 1347–1350.

Heslop-Harrison, J. (1991) The molecular cytogenetics of plants. *J. Cell Sci.*, **100**, 15–22.

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

Jiao, Y., Peluso, P., Shi, J., et al. (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524-527.

Jühling, F., Mörl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Pütz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.

Jurka, J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci.*, **94**, 1872–1877.

Kajikawa, M. and Okada, N. (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell*, **111**, 433–444.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., **30**, 3059–3066.

Kramerov, D.A. and Vassetzky, N.S. (2005) Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, **247**, 165–221.

Lagemaat, L.N. Van De, Gagnier, L., Medstrand, P. and Mager, D.L. (2005) Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res.*, **15**, 1243–1249.

Lenoir, A., Cournoyer, B., Warwick, S., Picard, G. and Deragon, J.M. (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. *Mol. Biol. Evol.*, **14**, 934–941.

Lenoir, A., Lavie, L., Prieto, J.L., Goubely, C., Coté, J.C., Pélissier, T. and Deragon, J.M. (2001) The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **18**, 2315–2322.

Lenoir, A., Pélissier, T., Bousquet-Antonelli, C. and Deragon, J.M. (2005) Comparative evolution history of SINEs in *Arabidopsis thaliana* and *Brassica oleracea*: evidence for a high rate of SINE loss. *Cytogenet. Genome Res.*, **110**, 441–447.

Lisch, D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.*, **14**, 49–61.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.

Luchetti, A. and Mantovani, B. (2013) Conserved domains and SINE diversity during animal evolution. *Genomics*, **102**, 296–300.

Mascher, M., Gundlach, H., Himmelbach, A., et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Mayer, K.F.X., Waugh, R., Langridge, P., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

Negroni, M. and Buc, H. (2001) Retroviral recombination: what drives the switch? *Nat. Rev. Mol. Cell Biol.,* **2**, 151–155.

Nishihara, H., Smit, A.F.A. and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.*, **16**, 864–874.

Nishihara, H., Terai, Y. and Okada, N. (2002) Characterization of novel Alu- and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. *Mol. Biol. Evol.*, **19**, 1964–1972.

Odom, G.L., Robichaux, J.L. and Deininger, P.L. (2004) Predicting mammalian SINE subfamily activity from A-tail length. *Mol. Biol. Evol.*, **21**, 2140–2148.

Okada, N. and Hamada, M. (1997) The 3′ ends of tRNA-derived SINEs originated from the 3’ ends of LINEs: a new example from the bovine genome. *J. Mol. Evol.*, **44**, 52–56.

Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. (1997) SINEs and LINEs share common 3’ sequences: a review. *Gene*, **205**, 229–243.

Ostertag, E.M. and Kazazian H.H., J. (2001) Twin priming: a proposed mechanism for the creation

of inversions in L1 retrotransposition. *Genome Res.*, **11**, 2059–2065.

Price, A.L., Eskin, E., Pevzner, P.A., Price, A.L., Eskin, E. and Pevzner, P.A. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.*, **14**, 2245–2252.

Roy, A.M., Carroll, M.L., Nguyen, S. V., Salem, A.H., Oldridge, M., Wilkie, A.O.M., Batzer, M.A. and Deininger, P.L. (2000) Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.*, **10**, 1485–1495.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.*, **81**, 8014–8018.

Schaack, S., Gilbert, C. and Feschotte, C. (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol.*, **25**, 537–546.

Schmidt, T., Schwarzacher, T. and Heslop-Harrison, J.S. (1994) Physical mapping of rRNA genes by fluorescent *in situ* hybridization and structural analysis of 5S rRNA genes and intergenic spacer sequences in sugar beet (*Beta vulgaris*). *Theor. Appl. Genet.*, **88**, 629–636.

Schmitz, J. (2012) SINEs as driving forces in genome evolution. *Repetitive DNA*, **7**, 92–107.

Schmitz, J. and Zischler, H. (2003) A novel family of tRNA-derived SINEs in the colugo and two new retrotransposable markers separating dermopterans from primates. *Mol. Phylogenet. Evol.*, **28**, 341–349.

Schnable, P.S., Ware, D., Fulton, R.S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schramm, L. and Hernandez, N. (2002) Recruitment of RNA polymerase III to its target promoters. *Genes Dev.*, **16**, 2593–2620.

Schwichtenberg, K., Wenke, T., Zakrzewski, F., Seibt, K.M., Minoche, A., Dohm, J.C., Weisshaar, B., Himmelbauer, H. and Schmidt, T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229–244.

Seibt, K.M., Wenke, T., Muders, K., Truberg, B. and Schmidt, T. (2016) Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. *Plant J.*, **86**, 268–285.

Takahashi, K. and Okada, N. (2002) Mosaic structure and retropositional dynamics during evolution

of subfamilies of short interspersed elements in African cichlids. *Mol. Biol. Evol.*, **19**, 1303–1312.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.

Tsuchimoto, S., Hirao, Y., Ohtsubo, E. and Ohtsubo, H. (2008) New SINE families from rice, OsSN, with poly(A) at the 3' ends. *Genes Genet. Syst.*, **83**, 227–236.

Ullu, E. and Tschudi, C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312**, 171–172.

Umeda, M., Ohtsubo, H. and Ohtsubo, E. (1991) Diversification of the rice *Waxy* gene by insertion of mobile DNA elements into introns. *Jpn. J. Genet.*, **66**, 569–586.

Vogel, J.P., Garvin, D.F., Mockler, T.C., et al. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Vu, G.T.H., Schmutzer, T., Bull, F., et al. (2015) Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome*, **8**, doi:10.3835/plantgenome2015.04.0021.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Wenke, T., Holtgräwe, D., Horn, A. V., Weisshaar, B. and Schmidt, T. (2009) An abundant and heavily truncated non-LTR retrotransposon (LINE) family in *Beta vulgaris*. *Plant Mol. Biol.*, **71**, 585–597.

Xu, J.-H., Osawa, I., Tsuchimoto, S., Ohtsubo, E. and Ohtsubo, H. (2005) Two new SINE elements, p-SINE2 and p-SINE3, from rice. *Genes Genet. Syst.*, **80**, 161–171.

Yadav, V.P., Mandal, P.K., Bhattacharya, A. and Bhattacharya, S. (2012) Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*. *Nat. Commun.*, **3**, 854.

Yagi, E., Akita, T. and Kawahara, T. (2011) A novel Au SINE sequence found in a gymnosperm. *Genes Genet. Syst.*, **86**, 19–25.

Yasui, Y., Nasuda, S., Matsuoka, Y. and Kawahara, T. (2001) The Au family, a novel short interspersed element (SINE) from *Aegilops umbellulata*. *Theor. Appl. Genet.*, **102**, 463–470.

Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. and Machida, Y. (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits

sequence homology to specific tRNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 6562–6566.

Zhang, X. and Wessler, S.R. (2005) BoS: a large and diverse family of short interspersed elements (SINEs) in *Brassica oleracea*. *J. Mol. Evol.*, **60**, 677–687.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.

Ziętkiewicz, E. and Labuda, D. (1996) Mosaic evolution of rodent B1 elements. *J. Mol. Evol.*, **42**, 66–72.

Zingler, N., Willhoeft, U., Brose, H., Schoder, V., Jahns, T., Hanschmann, K.O., Morrish, T.A. and Schumann, G.G. (2005) Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res.*, **15**, 780–789.

2.3 Comparative analysis of SINEs in Salicaceae species reveals 3' end diversification in many families

This study was resubmitted to 'The Plant Journal' after minor revisions on 5th of September 2019:

Kögler, A., Seibt, K. M., Heitkam, T., Morgenstern, K., Reiche, B., Brückner, M., Wolf, H., Krabel, D., Schmidt, T. (2019) Comparative analysis of short interspersed nuclear elements (SINEs) in Salicaceae species reveals 3' end diversification in many families.

Introduction

Plant genomes consist mainly of repetitive DNA such as tandemly arranged sequences (satellite DNA, telomers, rRNA genes) or dispersed transposable elements (TEs), constituting up to 80 % of the nuclear DNA of higher plants (Feschotte *et al.*, 2002; Baucom *et al.*, 2009; Oliver *et al.*, 2013).

Short interspersed nuclear elements (SINEs) belong to class I TEs (retrotransposons), propagate by a copy-and-paste mechanism and are non-coding and highly heterogeneous. SINE lengths range from 80 bp to 350 bp (Deragon and Zhang, 2006; Wenke *et al.*, 2011). Typically, plant SINEs are characterized by a composite structure: while the 5' SINE region is derived from tRNA genes providing the internal box A and box B promoter motifs for transcription by RNA polymerase III (pol III), the origin of the 3' SINE region is often unknown and highly variable.

The retrotransposition of SINE families depends on autonomous long interspersed nuclear elements (LINEs). The LINE reverse transcriptase creates new copies by reverse transcription of the SINE mRNA, starting at the SINE 3' tail sequence, usually made of adenine or thymine stretches (Ohshima and Okada, 2005; Dewannieux and Heidmann, 2005). During integration into the genome, target site duplications (TSDs) are created (Luan *et al.*, 1993; Ostertag and Kazazian, 2001). These TSDs are unique for each SINE as they reflect the flanking genomic regions at the integration site.

SINEs are widespread in angiosperm and gymnosperm genomes (Yagi *et al.*, 2011; Wenke *et al.*, 2011). However, their distribution often does not follow the phylogenetic relationships between species. It is suggested that SINEs, integrated and conserved in 'safe havens' (e.g. intronic regions), can be re-activated and propagated after long time of persistence, resulting in a patchy distribution pattern between different species (Schwichtenberg *et al.*, 2016; Fawcett and Innan, 2016). However, the mechanisms promoting the activation of these copies are largely unknown (Johnson and

Brookfield, 2006) and transpositional activity is not constant in scale during species evolution (Deininger and Batzer, 1995; Seibt *et al.*, 2016; Kögler *et al.*, 2017).

The SINE population of a genome often consists of families and subfamilies (Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016; Kögler *et al.*, 2017), which are structurally related, demonstrating that SINEs reshuffle during retrotransposition or recombine through nested integrations (Jurka *et al.*, 2005; Yadav *et al.*, 2012; Kögler *et al.*, 2017).

The Salicaceae are a plant family of woody plants and shrubs whose crown-group consists of the two genera *Populus* (poplar) and *Salix* (willow). Species are mainly diploid and characterized by a chromosome number of $2n = 2x = 38$ (Blackburn and Harrison, 1924). Their small genomes (~ 500 Mb), fast growth, the ability for vegetative propagation, and high environmental stress tolerance make the genus *Populus* attractive as a model system for deciduous tree genomics (Tuskan *et al.*, 2006). The reference genome sequence of *Populus trichocarpa* (Tuskan *et al.*, 2006) provided a basis for the selection of poplar clones optimized for sustainable energy production (Ragauskas *et al.*, 2006; Sannigrahi *et al.*, 2010).

In order to understand the SINE evolution in deciduous tree species, we analyzed the Salicaceae SINE landscape consisting of eleven SINE families comprising 27,077 full-length copies with dispersed genomic distribution and occurrence predominantly in euchromatic chromosomal regions. We uncovered the structural discrepancy between conserved 5’ SINE start motifs and diversification of the 3’ ends and showed that the high turnover of differing 3’ end variants is associated with periods of intense SINE activity and has resulted in multiple SINE subpopulations.

Experimental procedures

Computational methods

We used reference genome sequences of Salicaceae species, available at the JGI genome portal, the PlantGenIE platform and the NCBI homepage (Tuskan *et al.*, 2006; Sundell *et al.*, 2015) (Table S1). The *de novo* SINE identification was conducted with the *SINE-Finder* tool (Wenke *et al.*, 2011) using the following modifications from standard parameters: size of overlap (1,000 bp), TSD score cutoff (5 bp), and direction of TSD search (both directions). SINE candidate sequences derived as output were clustered with *UCLUST* (Edgar, 2010) using a similarity threshold of 60 %. Clusters were manually evaluated to separate SINE-like sequences from false positives, which show sequence conservation over the complete cluster length. The SINE-like sequences were realigned by *MUSCLE* (Edgar, 2004) to uncover the family structure. SINE clusters were identified by the conserved position and distance of RNA polymerase III promoter boxes of plant SINE families (Kögler *et al.*, 2017) and the presence of a poly(A) tail and varying sequences of the flanking target site duplications (TSDs).

The number of full-length copies per genome was determined by *BLAST* (Altschul *et al.*, 1990) searches using the SINE family consensus sequences derived from the SINE clusters as query. Resulting *BLAST* hits were aligned with *Geneious* Pro 6.1.8 (Kearse *et al.*, 2012) and pairwise identity values to the consensus sequence were used to discard sequences which were too diverged (similarity below 60 %). Of the first six 5' nucleotides of a SINE, at least two nucleotides have to match with the consensus sequence to represent a full-length copy. Truncated copies are not included as their assignment to the respective SaliS families is hardly possible due to segmental sequence similarities among the SINEs (Figure 2). Initially, SINE subfamilies were detected visually by identification of distinctive clusters in the multiple sequence alignment of SINE family members. The subfamily assignment has been verified by two additional approaches. First, the arrangement of representative SINE copies in the dendrogram was examined, constructed by *MEGA5* software (Tamura *et al.*, 2011) (neighbor-joining distance method and maximum composite likelihood nucleotide model) based on a *MUSCLE* multiple sequence alignment of 20 full-length copies for each putative subfamily with highest similarity to the species-specific consensus sequence. Second, a comparison of subfamily consensus sequences was conducted. The species-specific consensus sequences (Table S2) were

derived from *MUSCLE* alignments of all full-length copies, based on the most common bases, for the species with highest abundance. For subfamily definition, similarities had to range between at least 60 % and a maximum identity of 85 %.

The similarity of all SINE family members to the species-specific consensus sequence was calculated by pairwise comparisons using *Geneious* Pro 6.1.8 (Table S5, S6). For SINE families consisting of at least ten full-length copies, the percentage distances were represented by histograms containing similarity intervals (Figure 3, Figure S1). Tail sequences and TSDs were evaluated and statistically analyzed according to Kögler *et al.* (2017). Here, a sample of 20 SINEs were selected for each subfamily based on highest similarity to the species-specific consensus sequence. Normality of the data was inspected using the Shapiro-Wilk test and the correlation was examined using Spearman's rank correlation in *R* (R Core Team, 2017). The respective figures were generated using ggplot2 package in *R* (Wickham, 2016; R Core Team, 2017). For the investigation of the highly heterogeneous 3' ends, the different variants had to occur in at least 2 % of the full-length copies of the SINE family in the species analyzed.

To detect putative founding tRNA genes for the Salicaceae SINE families, the consensus sequences were used as queries for *BLAST* searches against the Viridiplantae tRNA gene database (Jühling *et al.*, 2009). The sequence logos for the conserved box A and box B motifs of the tRNA-derived promoter among Salicaceae SINE families and subfamilies were calculated using *Geneious* Pro 6.1.8 based on the most common bases.

To investigate the SINE association with genes, the positions of all full-length SINE copies on the *Populus trichocarpa* assembly (NCBI: GCA_000002775.3) were determined by exact string matching. Only SINEs located on the 19 pseudochromosomes (CM009290.1 to CM009308.1) were considered. Ambiguous copies were discarded. SINE positions were compared with gene coordinates from the genomic annotation file (NCBI: GCA_000002775.3) as previously described (Seibt *et al.*, 2016). In brief, we correlated genic SINEs with exon and CDS annotations to determine whether they are located in coding sequences (exon and CDS), untranslated regions (UTR; exon and not CDS) or introns (not exon and not CDS). For intergenic SINEs, the distance to the closest neighboring gene was calculated. Figures were prepared using ggplot2 package in *R* (Wickham, 2016; R Core Team,

2017). The genome portion of SINEs, genes and CDS was determined from the GFF annotation files using *R* libraries rtracklayer (Lawrence *et al.*, 2009) and GenomicRanges (Lawrence *et al.*, 2013). To account for overlapping annotations, intervals were preprocessed using the reduce function of the GenomicRanges package with the parameter ignore.strand=True.

Plant material and DNA isolation

Plants of *P. trichocarpa* (cultivar 'Weser 6') were obtained from the Staatsbetrieb Sachsenforst Graupa (Germany, www.sbs.sachsen.de). Fresh cuttings from *P. trichocarpa* were incubated in water to obtain roots for fluorescent *in situ* hybridization (FISH) analysis. Genomic DNA was extracted from leaf tissue using the SDS-based protocol according to Verbylaite *et al.* (2010).

Fluorescent *in situ* hybridization

Root tips of *P. trichocarpa* were incubated in 2 mM 8-hydroxyquinoline for 3 hours to accumulate metaphase chromosomes, followed by fixation in methanol/acetic acid (3:1). Chromosome preparation was conducted according to the following procedure:

After washing in water root tip meristems were incubated for 50 minutes at 37 °C with an enzyme solution containing 2.5 % (w/v) cellulase Onozuka R 10 (Serva), 2.5 % (v/v) pectinase from *Aspergillus niger* (Sigma), 1.0 % cytohelicase from *Helix pomatia* (Sigma), and 2.5 % pectolyase from *Aspergillus japonicus* (Sigma) in citrate buffer (4 mm citric acid, 6 mm natrium citrate, pH 4.5).

Root tips were transferred onto a pre-cleaned glass slide, treated with 45 % acetic acid, macerated and incubated at 50 °C for 1 minute. Subsequently, the nuclei suspension was spread over the glass slide. Hybridization probes of the SINE family SaliS-I were amplified from genomic DNA with specific primers (Table S5), cloned and labeled by PCR with biotin-11-dUTP (Roche). The probe sequence corresponds to a region shared by four SaliS families in *P. trichocarpa* with the highest similarity to SaliS-I (Figure S3). Similarity values were calculated in *Geneious* Pro 6.1.8 (Kearse *et al.*, 2012). The probe p18S for the 18S-5.8S-25S rRNA genes was derived from sugar beet (Paesold *et al.*, 2012) and labeled by nick translation with digoxygenin-dUTP and detected by antidigoxigenin–fluorescein isothiocyanate. Probes were hybridized *in situ* to mitotic chromosome spreads (washing stringency of

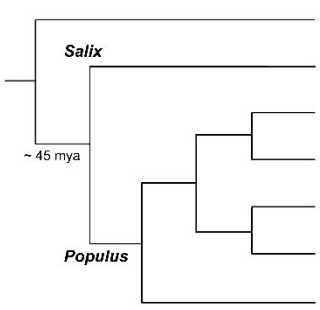
79 %), counterstained with DAPI (4′,6′-diamidino-2-phenylindole) (Heslop-Harrison, 1991). Fluorescence microscopy was conducted with a Zeiss Axioplan2 Imaging fluorescent microscope using filters 02 (DAPI) and 15 (Cy3). Images were acquired with the Applied Spectral Imaging v. 3.3 software coupled with the high-resolution CCD camera ASI BV300-20A. Image optimization only used functions of the Adobe Photoshop CS5 software affecting the whole image equally.

Results

Distribution and abundance of 20 SINE families and subfamilies in Salicaceae genomes

Six currently available poplar and willow genome sequences, namely *Populus deltoides*, *Populus euphratica*, *Populus tremula*, *Populus tremuloides*, *Populus trichocarpa*, and *Salix purpurea*, were compiled into a 2.4 Gb sequence data set (Table S1). They served as a basis for the identification of SINEs in the Salicaceae using the bioinformatic tool *SINE-Finder* (Wenke *et al.*, 2011). We detected eleven major SINE families, with four of them diverged into multiple subfamilies (Figure 1, Table S2). Together, the SINE families comprise 27,077 full-length copies in the six species. We previously identified five Salicaceae SINE families in an early *P. trichocarpa* assembly (Wenke *et al.*, 2011). Out of those, SaliS-I, SaliS-II, and SaliS-IV.2 are included in the SINEBase and RepBase databases as PTr-1, PTr-2, and PTr-3 (Jurka, 2010, Vassetzky and Kramerov, 2013; Bao *et al.*, 2015). Six SINE families have not been described yet and were designated accordingly as SaliS-VI to SaliS-XI. We now provide comprehensive details for the 20 SaliS families and subfamilies concerning their abundance, species distribution, similarity, and structure.

SaliS-I, SaliS-II and SaliS-III.1 populate all analyzed poplar species and willow with high copy numbers and therefore probably represent the most successful non-LTR retrotransposons of the SINE-type in these genomes (Figure 1). In general, we observed multiple examples of a patchy, mosaic-like SaliS distribution among the species tested (Figure 1). In some cases, SINE families are absent in a single genome, but present in closely related species, contradictory to the unidirectional propagation of SINEs, suggested to follow phylogenetic relationships (e.g. SaliS-IV.1 and SaliS-VII.2 lacking in *P. deltoides* and *P. tremula*, respectively). Others are specific for groups of closely related poplars only, such as SaliS-IV.3 and SaliS-VI.1 in *P. tremula* and *P. tremuloides*, respectively, and SaliS-IV.2 in *P. deltoides* and *P. trichocarpa*. Eight families and subfamilies (SaliS-III.2, SaliS-III.3, SaliS-VI.2, SaliS-VII.3, SaliS-VII.4, SaliS-IX, SaliS-X and SaliS-XI) are present only in *S. purpurea*, while SaliS-VI.3 occurs exclusively in *P. euphratica* (Figure 1).



	SaliS-I	SaliS-II	SaliS-III.1	SaliS-III.2	SaliS-III.3	SaliS-IV.1	SaliS-IV.2	SaliS-IV.3	SaliS-V	SaliS-VI.1	SaliS-VI.2	SaliS-VI.3	SaliS-VII.1	SaliS-VII.2	SaliS-VII.3	SaliS-VII.4	SaliS-VIII	SaliS-IX	SaliS-X	SaliS-XI	Total	Families and subfamilies
Bennettiodendron leprosipes																						
Idesia polycarpa																						
Salix purpurea	903	125	1830	281	59	851					320		26	394	23	10	64	175	7	16	5048	15
Populus tremula	1174	1922	264			902		33		172			18				1				4486	8
Populus tremuloides	1303	1548	260			677		24	1	198			8	3							4022	9
Populus deltoides	1136	1294	513				430		18				40	6							3437	7
Populus trichocarpa	2070	1333	292			402	92		42				30	1							4262	8
Populus euphratica	1350	391	152			3741						152									5786	5
Total	7936	6613	3311	281	59	6573	522	57	61	370	320	152	122	404	23	10	65	175	7	16	27077	**20**

Figure 1. Phylogenetic distribution of 20 Salicaceae SINE families and subfamilies. The distribution of SINE families and subfamilies is shown for each Salicaceae species (rows) and for each SaliS family and subfamily (columns). Numbers of full-length copies per SINE family (below), total copy number and number of families per species (right) are given. The average similarity of SaliS copies to the consensus sequence of the family (intraspecific diversity) is indicated by the colors red (60 % – 73 %), yellow (74 % – 87 %), and blue (88 % – 100 %). SINE families occurring in a species with only a single copy are excluded from the analysis. Species relationships and divergence times of the phylogenetic scheme (left) are modified from Liu *et al.* (2016).

The total number of SINEs in the analyzed species ranges from 3,437 in *P. deltoides* up to 5,786 in *P. euphratica*. The desert poplar *P. euphratica* contains also the highest number of full-length copies of a single SINE family (3,741 SaliS-IV.1 copies) reported for plant SINEs so far. The purple willow *S. purpurea,* a representative of the genus *Salix,* shows the largest number of different SINE families and subfamilies (n = 15, Figure 1).

The SINE families and subfamilies identified range in length from 158 bp (SaliS-VII.1 in *P. deltoides*) to 268 bp (SaliS-V in *P. trichocarpa*, Figure 2). SINEs are frequently subject to interelement reshuffling (Kögler *et al.*, 2017) accompanied by ongoing amplification and diversification. This resulted in SINE families sharing parts of their sequence, as we identified in 18 of the 20 SaliS families and subfamilies regions spanning 51 bp to 188 bp with a similarity of 75 % to 97 % (examples shown in Figure 2). These shared regions form three groups depending on the position within the SINEs. Structurally related 3' regions are found in SaliS-I, SaliS-II, SaliS-IV, SaliS-V, and SaliS-VI (Figure 2, orange). Moreover, the shared 135 bp sequence of the subfamilies SaliS-IV.1 and SaliS-IV.2 covers the whole SINE 3' region and even extends into the 5' SINE region. The second group comprises the three SINE families SaliS-III, SaliS-X, and SaliS-XI (Figure 2, green), whose internal regions most likely have the same origin. The third group consists of the subfamilies SaliS-VII.1 to SaliS-VII.4 with 75 % similarity over 100 bp to 137 bp, respectively, starting at the 5' regions (Figure 2, ochre).

Contrasting with common SINEs possessing a single promoter, SaliS-V is a composite SINE dimer with two promoter regions (Figure 2, white boxes). Its 3' region shows 97 % identity to the related monomeric variant SaliS-IV.2 (Figure 2, blue and orange), whereas the 5' region probably originated from a yet unknown SINE. Out of the identified SaliS families, it is the only SaliS heterodimer. We did not detect internal remnants of 3' tails, and thus no evidence of a recent nested integration of 5' truncated copies into existing SINEs.

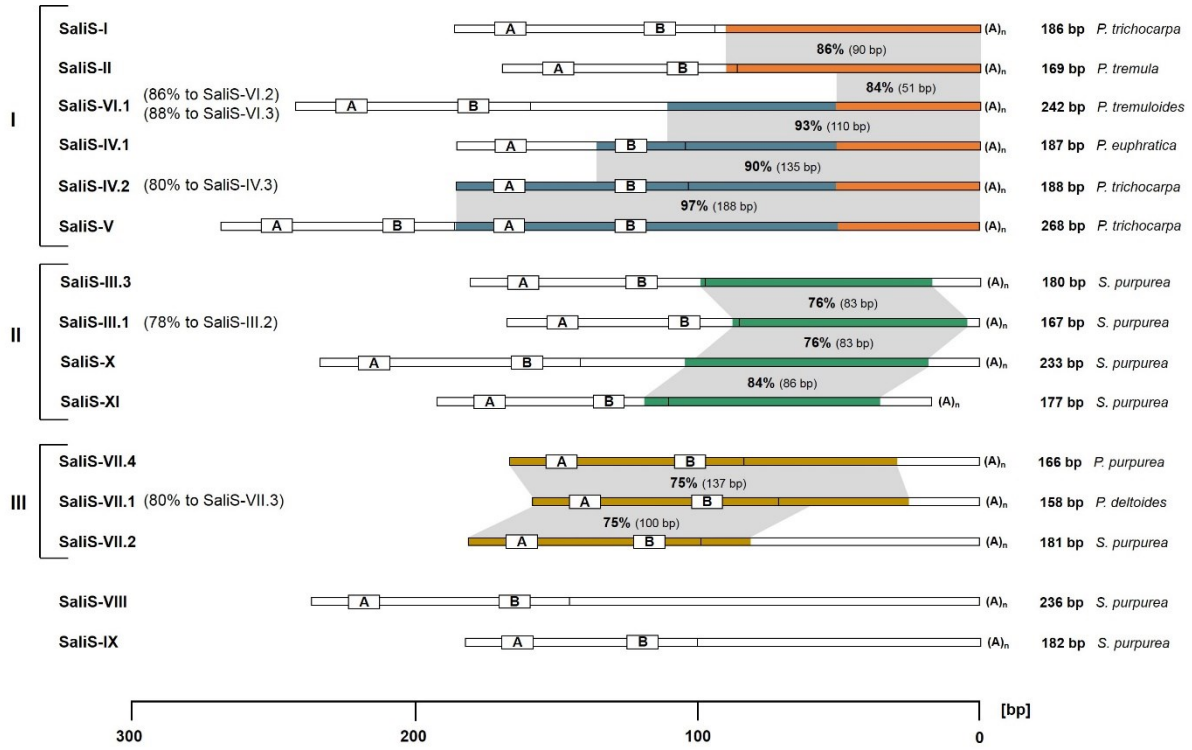


Figure 2. Structural relationships of Salicaceae SINE families and subfamilies. Sequence similarities (at least 75 % similarity over at least 50 bp) of three SINE groups (I-III) are shown by identical colors. Grey shadings illustrate related SINE regions, containing the percentage and length of sequence similarity. A black vertical line within the SINE marks the end of the tRNA-related SINE portion, 14 nucleotides after the box B motif (Deragon and Zhang, 2006). Promotor motifs (box A and B) of the SINE families are indicated by boxes. SINE families are represented with the consensus sequence of the species where the highest abundance was observed.

Evolutionary diversification into species-specific SINE landscapes

The distribution, abundance, and diversification of SINEs in plant genomes are determined by the transpositional activity of individual family members. We selected the four most abundant SaliS families with more than 3,000 full-length copies across the six species providing a robust dataset to analyze the SINE diversity within individual genomes and between different species (Figure 3). We calculated dendrograms to examine SINE diversification based on branching pattern and considered short branches lengths as indication for highly similar, presumably young SINE copies. The dendrograms contain 20 representative copies of each SaliS (sub)family with the highest similarity to the respective species-specific consensus sequence. We also examined the relative age of the family members by intervals of sequence similarity to the family consensus sequence, representing the putative founder SINE. We observed a characteristic intra- and interspecific SINE diversity and deduced two typical SINE differentiation patterns as follows:

(1) Undifferentiated and diversified SINEs in all species indicate long periods of inactivity

A striking example is the SaliS-II family which contains mostly diverse, evolutionarily ancient copies (long branches in the dendrogram, identities to the family consensus ranging from 60 % to 90 %) in the six species analyzed (Figure 3a). Recent amplificational bursts are absent and SaliS-II copies are randomly diverged, reflected by a star-like arrangement with SINEs from different species intermingled in the dendrogram (Figure 3a).

(2) Massive amplification of SINEs going along with differentiation into species-specific families and subfamilies. We exemplify this for the SINE families SaliS-I, SaliS-III, and SaliS-IV (Figure 3b-d).

In willow, SaliS-I was presumably inactive for a long time as only 24 of 903 copies resemble the species-specific consensus with more than 80 % sequence identity (Figure 3b, pink). The highly diverse willow SINEs (represented by long branches) are arranged separately from the majority of poplar SaliS-I copies in the dendrogram, demonstrating that recent SaliS-I differentiation only occurred in the *Populus* genus. Evolutionarily young SaliS-I copies are particularly numerous in *P. trichocarpa* and *P. euphratica* (histograms in Figure 3b, blue and orange). Of those, SaliS-I copies of the more distant *P. euphratica* are arranged on a separate branch, most likely reflecting a beginning differentiation to a new SINE subfamily (arrow 1 in Figure 3b), whereas SaliS-I copies from *P. trichocarpa* also have short branches and are intermingled with SINEs of other poplars (orange branches in Figure 3b).

In SaliS-III and IV, differentiation resulted in pronounced subfamilies visualized by the dendrograms (Figure 3c-d). Most striking, two out of three subfamilies (SaliS-III.2 and SaliS-III.3) are solely found in willow and only distantly related to subfamily SaliS-III.1 which contains both willow and poplar SINEs (Figure 3c). The analysis of the relative age indicates that SaliS-III.3 might presumably be still active, while the activity of SaliS-III.2 has ceased, as copies with more than 90 % similarity to consensus were not detected (Figure 3c, histograms).

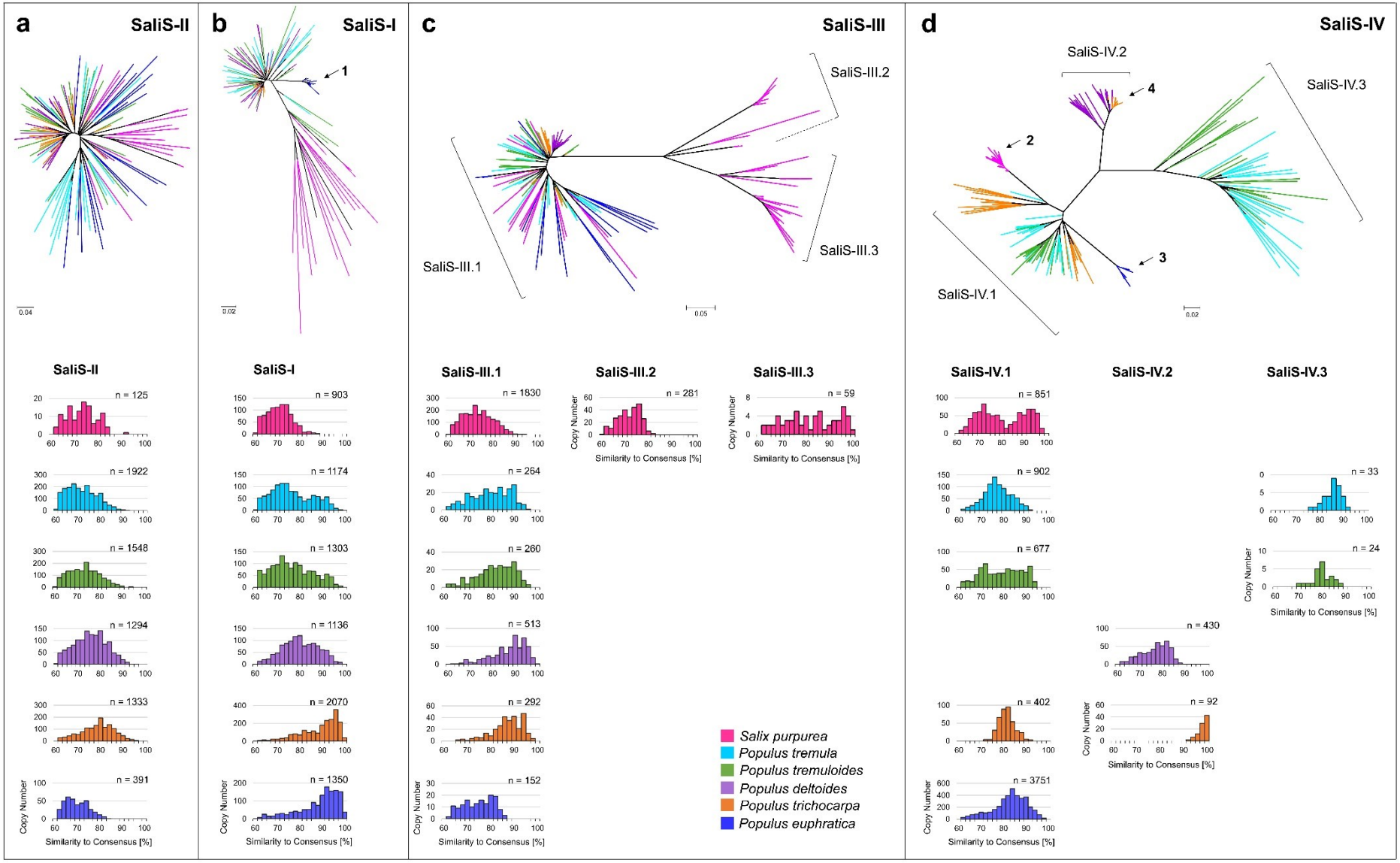

a
SaliS-II
0.04
b
SaliS-I
1
0.02
c
SaliS-III
SaliS-III.2
SaliS-III.3
SaliS-III.1
0.05
d
SaliS-IV
SaliS-IV.2
4
SaliS-IV.3
2
3
SaliS-IV.1
0.02
SaliS-II
n = 125
n = 1922
n = 1548
n = 1294
n = 1333
n = 391
SaliS-I
n = 903
n = 1174
n = 1303
n = 1136
n = 2070
n = 1350
SaliS-III.1
n = 1830
n = 264
n = 260
n = 513
n = 292
n = 152
SaliS-III.2
n = 281
SaliS-III.3
n = 59
SaliS-IV.1
n = 851
n = 902
n = 677
n = 402
n = 3751
SaliS-IV.2
n = 430
n = 92
SaliS-IV.3
n = 33
n = 24
Copy Number
Similarity to Consensus [%]
Salix purpurea
Populus tremula
Populus tremuloides
Populus deltoides
Populus trichocarpa
Populus euphratica

Figure 3. Inter- and intraspecific diversity of SaliS families and subfamilies. SINE families containing more than 3,000 members are represented by dendrograms with 20 representative copies of each Salicaceae species to demonstrate interspecific diversity. We distinguish between undifferentiated, highly diverged SINE populations (SaliS-II) (a) and examples of SINE differentiation (b-d), without subfamily formation (SaliS-I), species-specific subfamilies (SaliS-III) and the combination of both (SaliS-IV). Intraspecific diversity (below dendrograms) is indicated by similarity intervals reflecting recent transpositional activity of the SINE family members (n).

The three subfamilies of SaliS-IV are clearly distinct, but not restricted to a single species (Figure 3d). SaliS-IV.1 is present in five of the six analyzed species and shows different patterns of activity. While relatively continuous activity over a long period is detected in *P. tremuloides*, a short intense period of activity was found in *P. trichocarpa*. Gradually increasing and decreasing transpositional activity was observed for *P. tremula* and *P. euphratica.* Noteworthy, for *S. purpurea* two transposition maxima are observed. The SaliS-IV.1 copies of *P. euphratica* and *S. purpurea* evolved to species-specific variants and form clearly distinct groups in the dendrogram (arrows 2 and 3 in Figure 3d).

SaliS-IV.2 has full-length copies in *P. trichocarpa* and *P. deltoides*, only. In *P. trichocarpa* it represents an example for a recent activation of a SINE subfamily, as 43 out of 92 copies show 96 %- 100 % sequence identity to the SINE consensus (arrow 4, Figure 3d). However, these copies are still highly similar to those of *P. deltoides*, which are slightly more diverged (Figure 3d). Presumably, the sequence of a retrotransposition-competent SaliS-IV.2 copy remained conserved over a long period. Compared to the other two subfamilies, SaliS-IV.3 is more diverged and shows species-specificity for *P. tremula* and *P. tremuloides*.

Physical mapping of SINEs along poplar chromosomes and their association to genes

Plant SINEs are randomly distributed along all chromosomes, but often excluded from heterochromatic regions (Deragon and Zhang, 2006; Baucom *et al.*, 2009; Wenke *et al.*, 2011; Kögler *et al.*, 2017). Furthermore, some SINEs accumulate in the vicinity of genic regions (Ben-David *et al.*, 2013; Seibt *et al.*, 2016; Keidar *et al.*, 2018). We investigated the chromosomal distribution of SINEs exemplarily for SaliS-I in *P. trichocarpa* using fluorescent *in situ* hybridization (Figure 4) and the association of SINEs to euchromatic gene-rich regions by *in silico* mapping to *P. trichocarpa* pseudochromosomes (Figure 5).

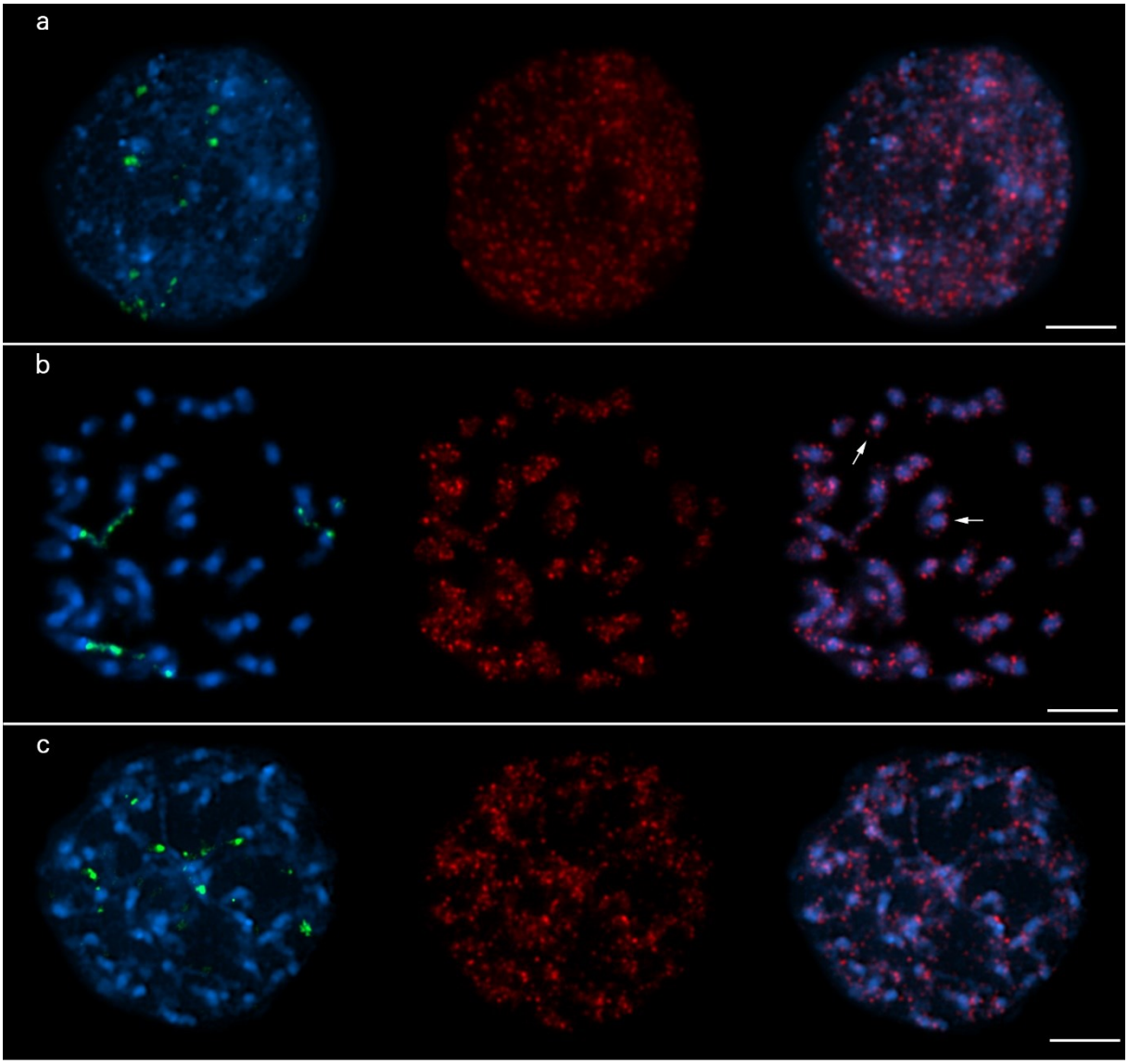


Figure 4. Physical mapping of SINEs on *P. trichocarpa* chromosomes by FISH. The DNA in *P. trichocarpa* chromosomes (blue) is stained with DAPI. Red signals at interphase (a), mitotic metaphase (b), and prometaphase (c) are sites of SaliS-I and SaliS-II hybridization, showing the dispersed chromosomal distribution. The arrow points to an example of signal doublets on both chromatides. Positions of the 18S-5.8S-25S rRNA genes is indicated by a green fluorescence. The scale bar corresponds to 5 μm.

The probe used for FISH (Table S3) was derived from the 3' region of SaliS-I to avoid cross-hybridization to promoter regions of tRNA genes, but also to enable the detection of structurally related SINE families (e.g. SaliS-II, Figure S3).

Interphase nuclei (Figure 4a), mitotic metaphase (Figure 4b) and prometaphase (Figure 4c) chromosomes of *P. trichocarpa* show an interspersed distribution of -SINEs and an accumulation in terminal regions, often visible on both chromatids (example arrowed in Figure 4b). In interphases, exclusion or depletion from strongly DAPI-stained heterochromatin was observed.

We were able to assign 3,870 of the 4,262 full-length SaliS copies identified in *P. trichocarpa* to the 19 pseudochromosomes (NCBI, GCF_000000955.4) of the reference sequence to analyze their physical relationship to genes: More than 95 % of SaliS copies (3687 of 3870 full-length SINEs, Table S4) are located in intergenic regions. Nevertheless, more than 25 % are within 1 kb distance to an annotated gene and the majority of copies is located in less than 5 kb distance (Figure 5). SaliS-VII.2 copies are not associated with genes.

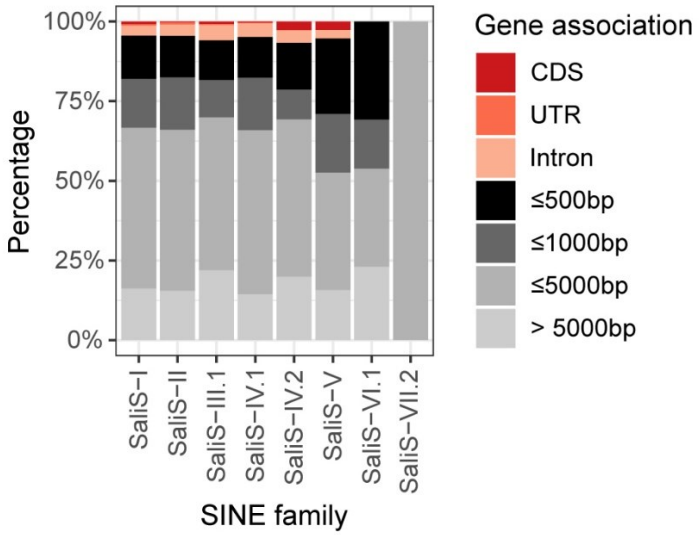


Figure 5. Location of genomic SaliS full-length copies in relation to the closest gene. Full-length SaliS copies were mapped to the *P. trichocarpa* pseudochromosomes to reveal their association with genes. The majority of SINEs is integrated in the 5000 bp 5' or 3' flanking region of genes (distance intervals in grey scales).

Conservation of crucial SINE regions

Apart from the conserved SINE family structure, some features uniquely define each copy, such as diagnostic point mutations, target site duplications (TSDs), and the length of the 3' tail.

The similarities of the SINE copies to the species-specific SINE family consensus sequence were averaged and are color-coded in Figure 1. SaliS-II in *P. euphratica* is the most diverged SINE family (69 % average similarity, Table S5), whereas SaliS-V and SaliS-VII to SaliS-X are examples for evolutionarily young SINE families (88 % - 100 % average similarity, Figure 1, Table S5).

We analyzed the average lengths of the TSDs and the 3' tails for each family in each species based on 20 representative copies (Table S6). The 3' tails of Salicaceae SINEs consist exclusively of adenine stretches. The average lengths of these poly(A) tails range between 8 bp (SaliS-I and SaliS-X in *S. purpurea*, SaliS-VII.1 in *P. trichocarpa*) and 21 bp (SaliS-IV.1 in *P. euphratica*) (Table S6). In

general, 48 of 52 inspected poly(A) tails vary between 8 bp and 14 bp (Figure S4a, Table S6). The maximum tail length observed was a 32 bp adenine stretch of a SaliS-I copy in *P. euphratica* (not shown).

During SINE integration into the host genome, a target site duplication of variable length is generated. Average lengths of the flanking TSDs are mainly between 10 bp and 14 bp (Figure S4a) and range from 9 bp (SaliS-II of *P. euphratica*) to 17 bp (SaliS-VII.2 in *P. tremuloides* and SaliS-IX in *P. tremula*) (Table S6). The maximum TSD length was found in a SaliS-VIII copy of *S. purpurea* consisting of 22 bp.

Figure 6 correlates the TSD length with the SINE similarity of 20 representative SINE copies. It shows the TSD lengths of all SaliS families, arranged by decreasing TSD length, and the similarity values following this decline.

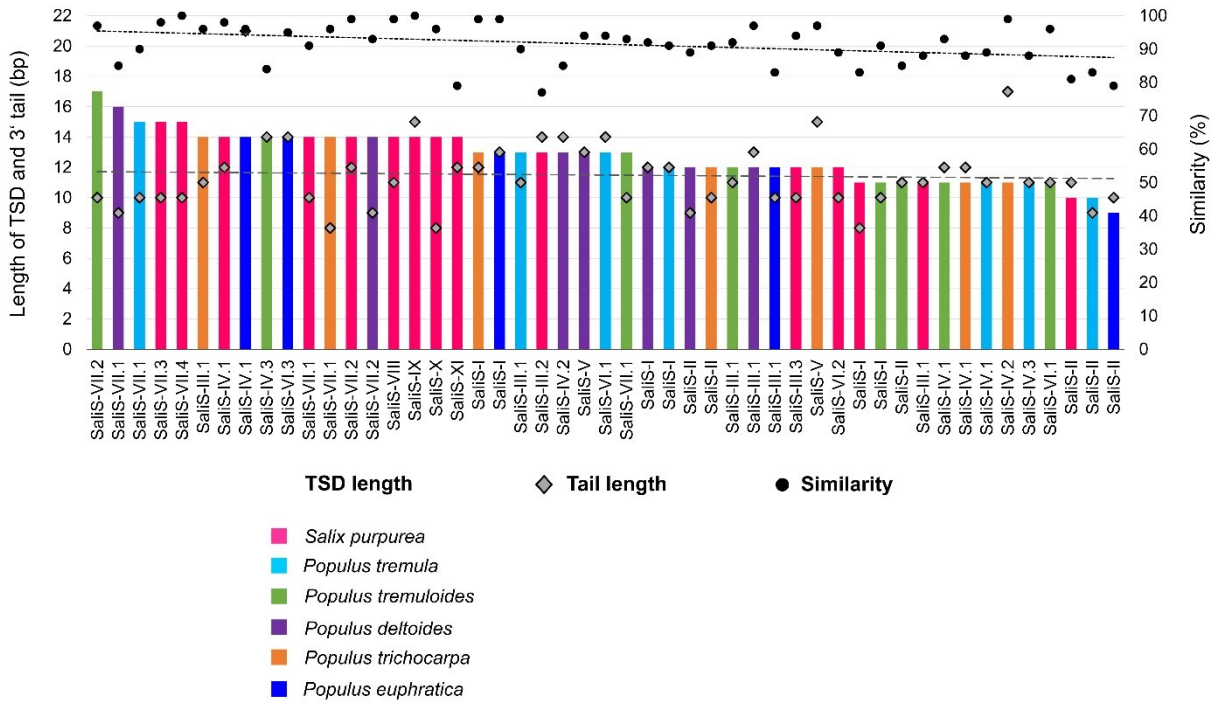


Figure 6. Correlation of average similarity and TSD length of Salicaceae SINEs. The average TSD lengths (bars) of SINE families and subfamilies of each Salicaceae species investigated are arranged by decreasing size and compared with the average length of 3' tails (grey diamonds) and the average similarity of SINE family members (black dots). Linear trend lines of tail length (y = -0.0094x + 11.724) and similarity (y = -0.1661x + 95.602) are indicated by a dashed line and a dot-dashed line, respectively. SaliS-V in *P. tremuloides*, SaliS-VII.2 in *P. trichocarpa*, and SaliS-VIII in *P. tremula* (one full-length copy only) are not included. Statistical tests are described in Figure S4.

Statistical tests revealed a significant positive correlation between TSD lengths and similarity ($p = 0.0017$ and $rho = 0.4362$, Figure S4b). According to Cohen (1992) the positive correlation is of medium effect size ($rho > 0.30$). As SINEs and the flanking TSDs accumulate mutations over time, the TSD will decay by point mutations or indels, until it is shortened and barely recognizable. The length of the poly(A) tails was neither correlated to SINE similarity nor to TSD lengths.

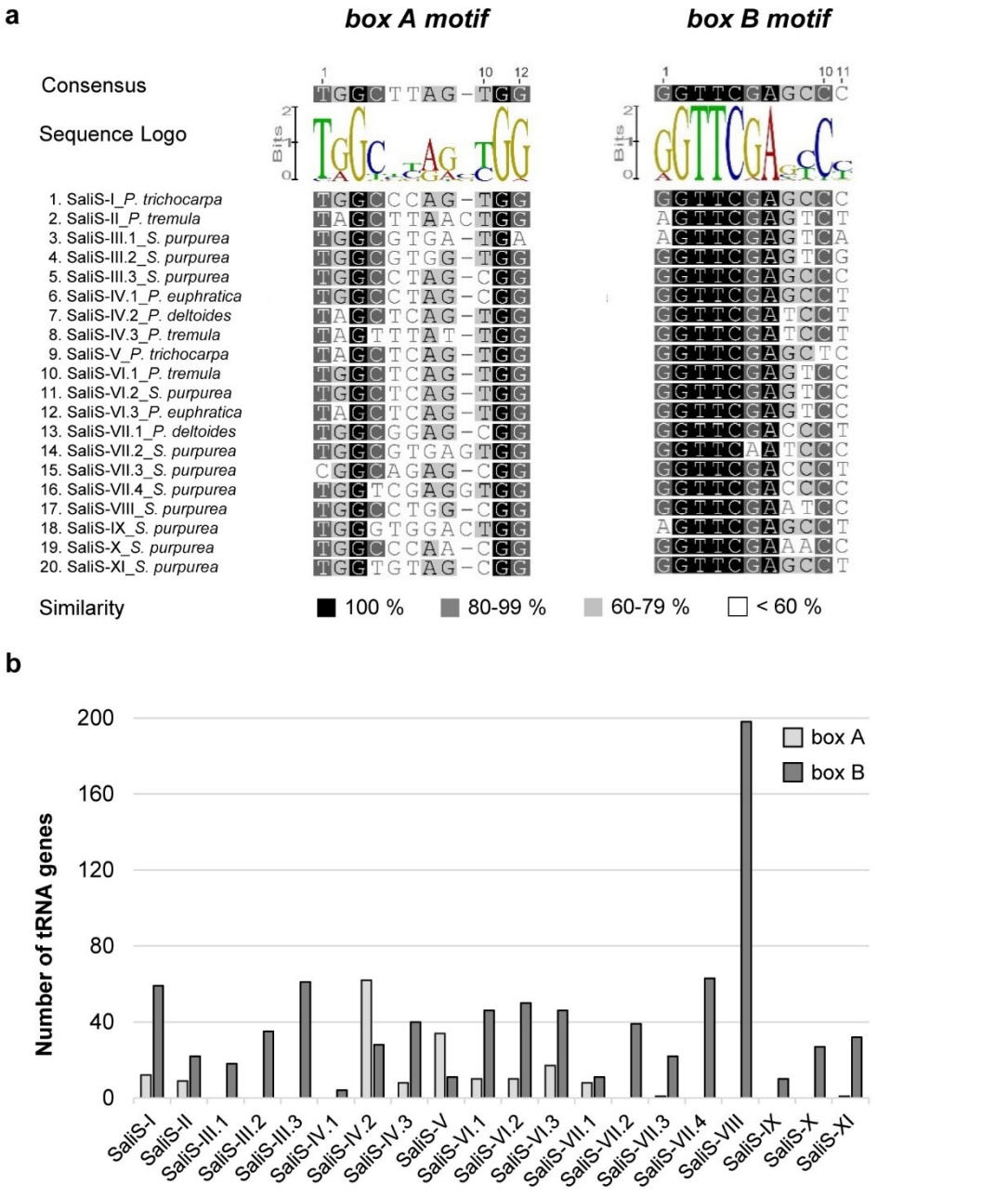


Figure 7. RNA polymerase III promotor motifs of SaliS families and subfamilies. (a) Conserved nucleotides of box A (11 bp - 12 bp) and box B (11 bp) are shown by consensus sequences based on the most common bases. (b) In total, 702 Viridiplantae tRNA genes were mapped to the SaliS promotor motifs. Matches are included in case of at least eight consistent nucleotides per box motif.

Crucial for the transcription of tRNA-derived SINEs is the RNA polymerase III promoter in the 5' region consisting of box A and box B. We analyzed the two 11 bp motifs in all SINE families and observed the following conserved promoter motifs within the Salicaceae: TGGCNNAGTGG for box A and GGTTCGAGCCN for box B (nucleotides with 100 % conservation underlined, nucleotides with less than 60 % conservation indicated by 'N'). The underlying SINE sequences and sequence logos are shown in Figure 7a. The two promoter boxes are mainly separated by 31 bp to 33 bp. Only SaliS-I, SaliS-VIII (42 bp each), and SaliS-X (43 bp) exhibit an enlarged distance between box A and B motif (not shown). Homology to a specific tRNA gene has not been observed, the conservation is restricted to the promoter box motifs required for transcription. Apparently, box B is more conserved than box A, which is consistent with the number of matches obtained by a *BLAST* search against 702 Viridiplantae tRNA genes (Jühling *et al.*, 2009). On average, 41 tRNA genes match to the box B of SaliS families, while only nine tRNA genes fit with box A. Out of the 20 SINE (sub)families, only two SINE families, SaliS-IV.2 and the dimeric SaliS-V, show more matches for box A than for box B, whereas nine SaliS families produced no box A tRNA match. For box B, a maximum of 198 matches with at least eight of eleven nucleotides identical to tRNA genes were obtained (SaliS-VIII, Figure 7b).

Conserved 5' start motifs of the SaliS families contrast with heterogeneous 3' ends

We next specifically analyzed the 5' and 3' ends as they are delimiting the SINEs from the flanking genomic neighborhood. The fine-scale comparison of SaliS families revealed two SINE groups based on their 5' start motifs consisting of the first ten nucleotides of the SaliS consensus sequences. We identified the prevalent motif ACCCANNNGG in twelve families and subfamilies (nucleotides with less than 60 % conservation indicated by 'N', Figure 8, Figure S2). The second group comprises SaliS-III.1 and four *Salix*-specific SINEs (SaliS-III.2, SaliS-III.3, SaliS-X, and SaliS-XI) starting with the conserved nucleotides GTCCCCGAGG (Figure 8, Figure S2). The three remaining SINE families show different 5' start motifs (SaliS-IV.1 and SaliS-IV.2 with GCAMTYRAGG; SaliS-II with AATTTTGAGG).

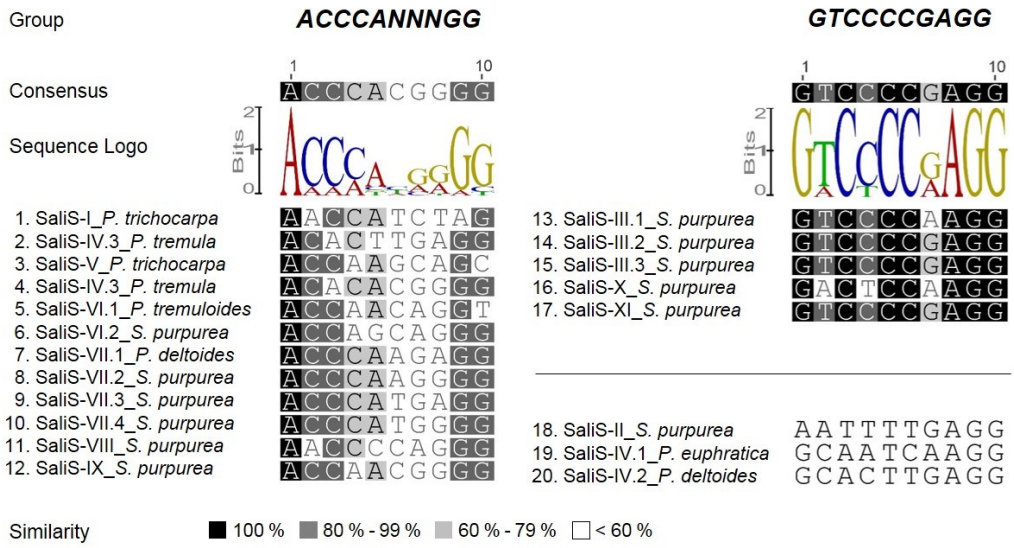


Figure 8. Groups of 5' start motifs of Salicaceae SINE families and subfamilies. Salicaceae SINE families and subfamilies fall into two different groups concerning the first six nucleotides of the 5' end. Only three SINE families (SaliS-II, SaliS-IV.1, SaliS-IV.2) show other 5' motifs and are listed separately. For each group, the first ten 5' nucleotides of the consensus sequence of all Salicaceae SINE families and subfamilies (species with highest abundance) are shown with the consensus sequence and the respective sequence logo.

As these 5' start motifs are conserved across SINE families, we deduced that they are a hallmark of SaliS families (Figure 8). In contrast, the 3' ends differ among the analyzed Salicaceae SINE (sub)families and species (Table S7). We observed that the family-specific sequence conservation is not directly connected to the poly(A) tail, but is separated by a variable A/T-rich region following the three terminal conserved 3' nucleotides of the SINE. We designated these three nucleotides 'terminal conserved triplet'. The sequence motifs spanning the terminal conserved triplets to the poly(A) tail are heterogeneous and form different fractions in the SaliS families within a species (Figure 9a, Table S7). For example, the terminal conserved triplet of SaliS-III.1 is "TCG", followed by three possible 3'end sequences, with a preference of the "AATC" 3' end (Figure 9b, yellow).

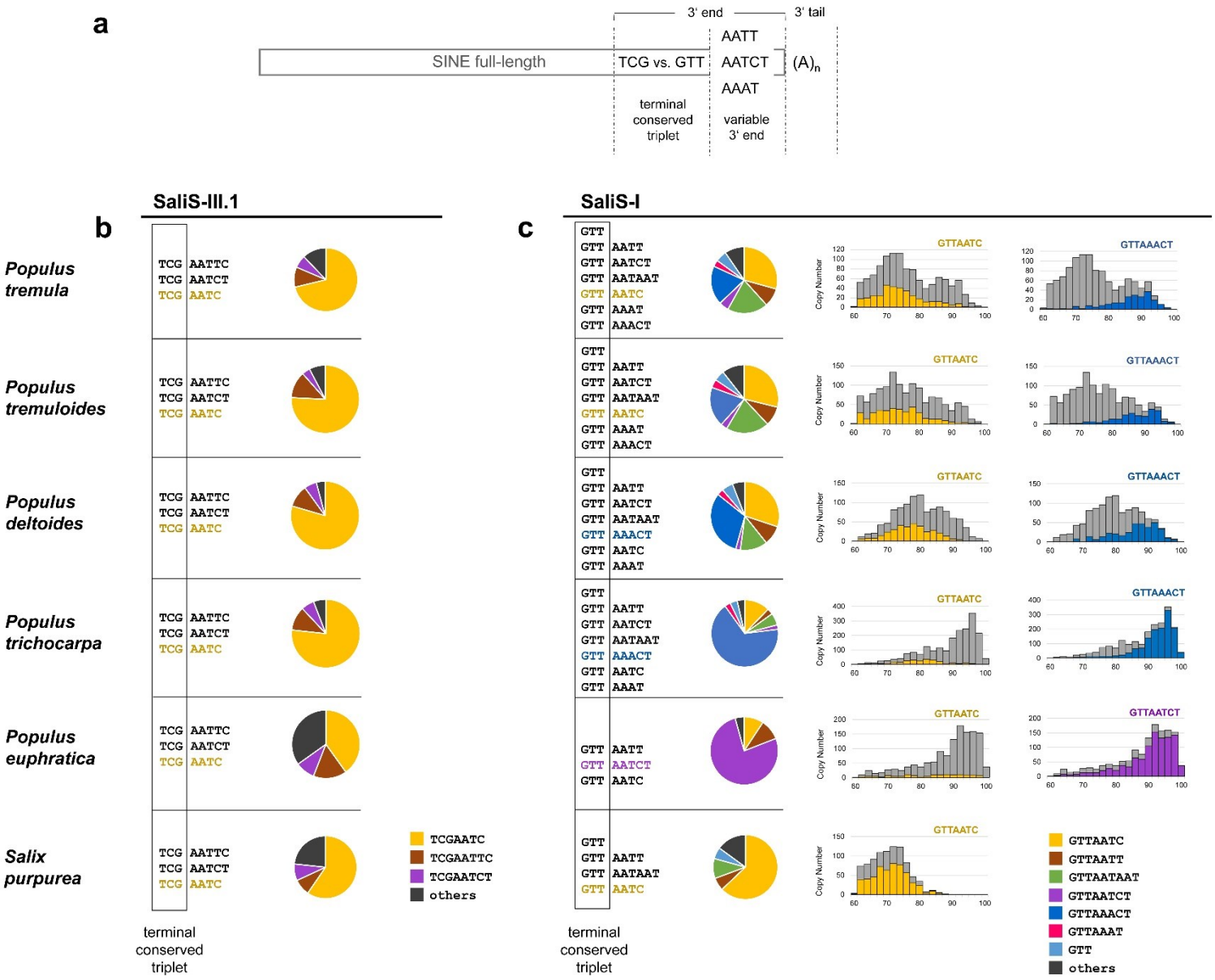

a
3' end
3' tail
AATT
SINE full-length
TCG vs. GTT
AATCT
(A)n
AAAT
terminal conserved triplet
variable 3' end
b
SaliS-III.1
Populus tremula
Populus tremuloides
Populus deltoides
Populus trichocarpa
Populus euphratica
Salix purpurea
TCG AATTC
TCG AATCT
TCG AATC
TCGAATC
TCGAATTC
TCGAATCT
others
terminal conserved triplet
c
SaliS-I
GTT
GTT AATT
GTT AATCT
GTT AATAAT
GTT AATC
GTT AAAT
GTT AAACT
GTTAATC
GTTAAACT
GTTAATCT
GTTAATT
GTTAATAAT
GTTAAAT
Copy Number
terminal conserved triplet

Figure 9. Comparative quantification of the 3' end motifs of SaliS-III.1 and SaliS-I. (a) The different 3' ends are composed of the terminal conserved triplet and different variable 3' end sequences, together forming the 3' end motif. The 3' end motifs of SaliS-III.1 (b) and SaliS-I (c) are listed for each Salicaceae species (left) and the most frequent 3' end is drawn in the respective color. Pie charts show their quantitative distribution. For poplar SaliS-I SINEs, activity profiles (see also Figure 3) indicate the age of copies of the major 3' end variants in corresponding colors. The portion of the more ancient 3' end AATC (yellow) decreased in case of more recently amplification spreading alternative 3' ends (blue, purple).

In some SINE families, the terminal triplets are conserved in all Salicaceae species, but frequently they also vary species-specifically (Table S7). For example, in SaliS-II, "GTT" is the main terminal conserved triplet in *S. purpurea*, *P. deltoides* and *P. trichocarpa*, while "ATT" is predominantly observed in *P. tremula*, *P. tremuloides* and *P. euphratica* (Table S7). Across all SaliS (sub)families, AATC represents the most frequent type of variable 3' end sequences of poplar SINEs, widespread in SaliS-I to SaliS-VI (Table S7). SINEs predominantly distributed in willow (Figure 1) are characterized by a higher sequence similarity (SaliS-VII to SaliS-XI) and show a reduced number of 3' end subpopulations (Table S7) with the most widespread variable 3' end sequence ACC.

In order to uncover the SINE 3' end evolution, we classified and quantified exemplarily the 3' ends in the SINE families SaliS-I and SaliS-III.1 which contain the conserved terminal triplets GTT and TCG, respectively, across all six genomes. We observed highly variable 3' end sequences (charts, Figure 9b, c).

In SaliS-III.1, three major variable 3' ends can be distinguished: Most SaliS-III.1 copies (2168 of 3311) of all analyzed species end with 5'-AATC-3', with moderate levels of variation in their frequency (yellow, Figure 9b). Almost identical fractions were observed for each of the two pairs of closely related poplars (*P. deltoides*/*P. trichocarpa* and *P. tremula*/*P. tremuloides*): In these species, the second most frequent 3' end (AATTC) accounts for 10 % - 13 % of all full-length copies (brown, Figure 9b), whereas the third 3' end (AATCT) includes between 4 % and 6 % of all SaliS-III.1 members (purple, Figure 9b). In *S. purpurea* and *P. euphratica* SaliS-III.1 is more diverged (Figure 3c) and accordingly diversified 3' ends were detected (category 'others', dark grey, Figure 9b).

In contrast to SaliS-III.1, the most frequently occurring SaliS-I 3' end varies across the six Salicaceae species and we detected three to seven distinct 3' end motifs (Figure 9c). Noteworthy, SaliS-I copies with the 3' end AATC are ancient and more diverged across all species (yellow fraction in activity

profiles; Figure 9c), while *Populus* SaliS-I copies ending with AAACT and AATCT are more homogeneous (blue and purple fraction in activity profiles; Figure 9c). The successive replacement of the ancient AATC end (yellow; Figure 9c) in some species is consistent with the SINE activity: in *P. trichocarpa* and *P. euphratica,* the SaliS-I family contains evolutionarily younger SINE copies, most of which represent the novel 3' ends AAACT (blue, Figure 9c) and AATCT (purple, Figure 9c). Taken together, the detailed analysis of SaliS terminal sequences revealed a conservation of the SINE 5' start across SINE families and species with two distinct sequence motifs. In contrast, within the SINE families the variability of the sequence preceding the poly(A) tail (variable 3' end) might be determined by the SINE amplification patterns.

Discussion

Massive SINE amplification during salicoid duplication in the *Populus-Salix* progenitor

We identified 27,077 full-length SINE copies in five poplar (*P. deltoides, P. euphratica, P. tremula, P. tremuloides, P. trichocarpa*) and one willow species (*S. purpurea*), falling into 20 SINE families and subfamilies detected with the *SINE-Finder* tool (Wenke *et al.*, 2011). All Salicaceae SINE families can be divided into two categories: the group of highly abundant SINE families, broadly distributed in all analyzed species (SaliS-I, SaliS-II, SaliS-III.1, SaliS-IV.1, and SaliS-VII.1), and the group of moderately abundant SINE families with a patchy distribution and an often more recent amplification (Figure 1, Figure S1).

Current molecular phylogenetic studies (Wang *et al.*, 2014; Lauron-Moreau *et al.*, 2015; Liu *et al.*, 2016) revealed that *Populus* and *Salix* are monophyletic sister genera comprising more than 450 willow species (Argus *et al.*, 2010) and 29 to 32 poplar species (Eckenwalder, 1996; Dickmann and Kuzovkina, 2014). Both genera show remnants of an ancient whole genome duplication ('salicoid duplication'), which occurred in a common ancestor approximately 65 million years ago (mya) (Tuskan *et al.*, 2006; Dai *et al.*, 2014). The resulting paleotetraploid progenitor was subject to intense genome reorganization leading to its diploidization (Dai *et al.*, 2014; Hou *et al.*, 2016). According to fossils, the divergence of the *Populus* and *Salix* lineages dates back from 60 to 65 mya (Collinson, 1992; Eckenwalder, 1996) up to 45 mya (Boucher *et al.*, 2003; Manchester *et al.*, 2006), respectively.

Whole genome duplications (WGDs) and polyploidizations are linked with TE activation and expansion and are responsible for genome reshaping (Wendel *et al.*, 2016; Vicient and Casacuberta, 2017). The ancient WGD might explain the amplification of six SINE families (SaliS-I, SaliS-II, SaliS-III.1, SaliS-IV.1, SaliS-VII.1, and SaliS-VII.2), which were probably present in the common ancestor of *Salix* and *Populus* (Figure 10).

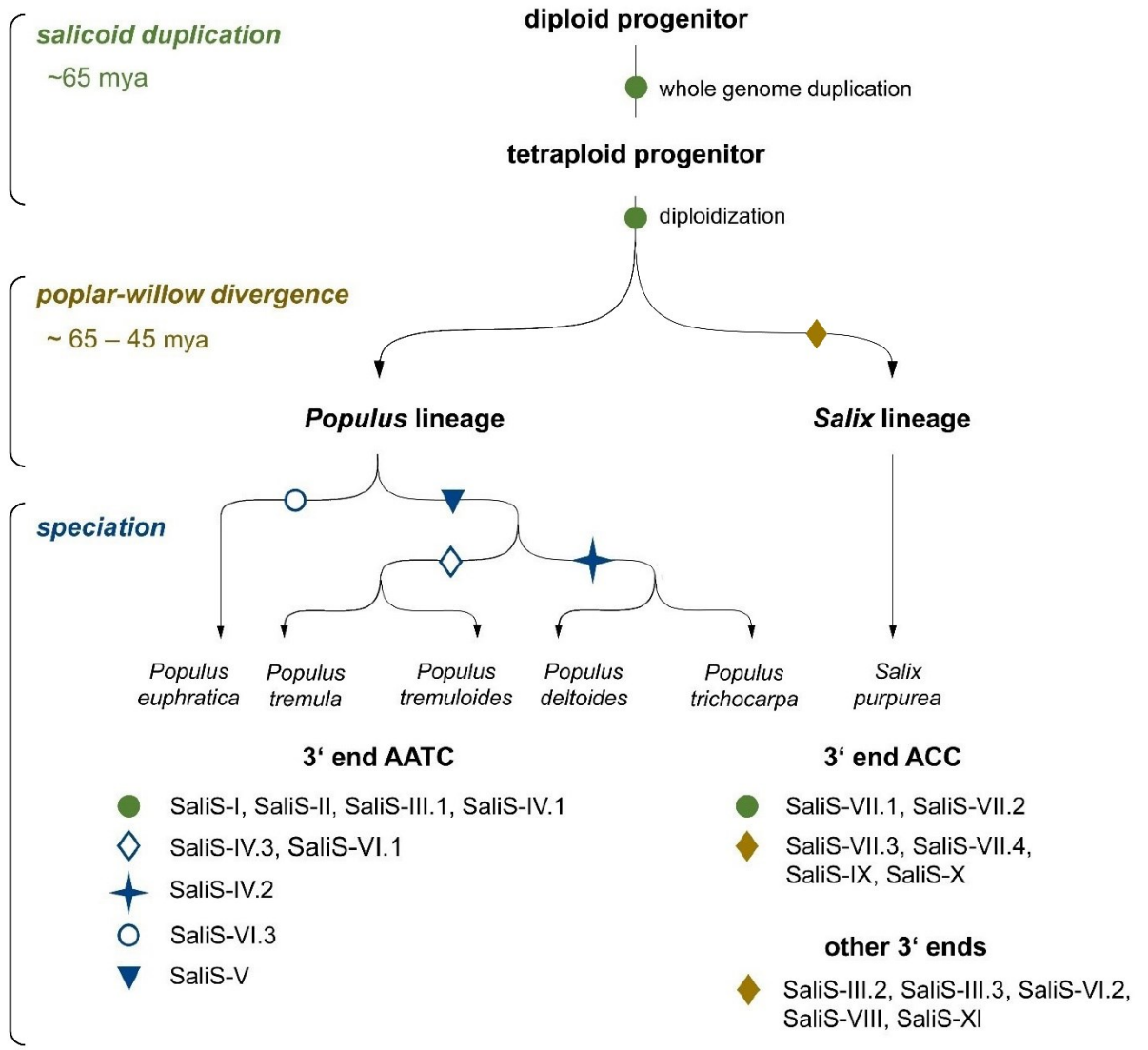


Figure 10. Evolutionary scenario for the SINE diversification during speciation within the Salicaceae. The genome rearrangements following the salicoid duplication probably caused a massive SINE expansion (green dots). SINE amplification continued also during divergence of the genera *Populus* and *Salix* (ochre diamonds) and during the speciation in the *Populus* lineage (blue family-specific symbols). Estimated SINE amplification (not emergence) is indicated by respective symbols. The time estimate of the salicoid duplication is taken from Tuskan *et al.* (2006). Time estimates for poplar-willow divergence are taken from Collinson (1992), Eckenwalder (1996), Boucher *et al.* (2003), and Manchester *et al.*(2006).

However, not all TE families show a similar response to genome rearrangements (Senerchia *et al.*, 2014). The high number and exclusive occurrence of SINE families in willow could be related to the 'poplar-to-willow' chromosome rearrangements in *Salix* species proposed by Dickmann and Kuzovkina (2014) and Hou *et al.* (2016). This probably gave rise to willow-specific SINEs (SaliS-III.2, SaliS-III.3, SaliS-VI.2, SaliS-VII.3 to SaliS-XI, Figure 10).

Interestingly, sequence similarities have been detected outside of the Salicaceae, as SaliS-VIII is similar to the Julia-SINEs from walnut and SolS-VI from Solanaceae plants, particularly pronounced in the 3' region (Wenke *et al.*, 2011; Wu *et al.*, 2012). We assume, that these SINEs evolved from a common ancestral SINE family and have been conserved in different plant families.

Diversification, differentiation and sequence reshuffling are main evolutionary processes for the generation of new SINE families

We observed structural relationships of SaliS families (Figure 2) with similarities between internal (SaliS-III, SaliS-X, SaliS-XI), 5' (SaliS-VII) or 3' regions (SaliS-I, SaliS-II, SaliS-IV, SaliS-V, SaliS-VI). Only two SINE families (SaliS-VIII and SaliS-IX) are not related to other Salicaceae SINEs identified in this study. The composite structures of SINEs, previously described for Poaceae SINE families (Kögler *et al.*, 2017), the TS SINE in tobacco (Wenke *et al.*, 2011) and the BoS SINEs in Brassicaceae (Deragon and Zhang, 2006), suggest that reshuffling by nested retrotransposition or recombination are the main evolutionary processes for the emergence of novel SINE families.

SaliS-V is a heterodimeric SINE family, presumably originating from retrotransposition of a SaliS-IV.2 copy into a yet unknown SINE. The 3' SINE region of the SaliS-V dimer resembles SaliS-IV.2 with 97 % sequence identity, which is also populating, similar to SaliS-V, only *P. deltoides* and *P. trichocarpa* genomes (Figure 1). We suggest a recent amplification of SaliS-V in the *P. deltoides*-*P. trichocarpa*-lineage, which is supported by the high average similarity of SaliS-V with 94 % and 95 %, respectively (Table S5). However, we also found a single SaliS-V copy in *P. tremuloides* highly similar (93 % each) but with discriminative point mutations relative to the consensus elements of the other poplar species (Figure S5). Therefore, the nested structure of SaliS-V most likely arose in an ancestor of poplars, has been preserved in *P. tremuloides* and amplified lineage-specifically in

P. deltoides and *P. trichocarpa* or in the common ancestor of these species (Figure 10). Recently, dimerization was also described for the homodimeric SINE PoaS-XIV from wheat, resembling two full-length copies of the same SINE subfamily (Kögler *et al.*, 2017). SINE dimerization is well documented in animal SINEs (Ullu and Tschudi, 1984; Feschotte *et al.*, 2001; Churakov *et al.*, 2005), including also combined tRNA- and 7SL-derived SINEs (Nishihara *et al.*, 2002).

In general, nested integration is common for retroelements (SanMiguel *et al.*, 1996; Levy *et al.*, 2009; Weber and Schmidt, 2009; Gao *et al.*, 2012) and the combination of different repeats creates new composite retroelements, e.g. the SVA in the human genome (Buzdin, 2004). The tendency to form clusters (J Jurka *et al.*, 2005) and their potential accumulation in or close to genes (Seibt *et al.*, 2016) results in a high SINE density, which increases the probability of nested SINE integration. Modular evolution and reshuffling has been observed in many transposable elements (Wollrab *et al.*, 2012; Smyshlyaev *et al.*, 2013) and might also result from illegitimate recombination, unequal homologous recombination (Katrien M Devos *et al.*, 2002; Ma *et al.*, 2005) or from a template switch of the reverse transcriptase (Marco and Marín, 2008; Du *et al.*, 2010; Yadav *et al.*, 2012).

Different rates of SINE divergence

We compared the diversity of Salicaceae SINE copies on the species and the SINE family level to gain insights into their evolution (Figure 3, Figure S1). Our data allowed us to detect undifferentiated and species-specific Salicaceae SINE populations resulting from both, retrotranspositional activity and diversification. Various patterns of SINE activity such as continuous retrotransposition or amplificational bursts are contrasting examples, which have also been described in other plant (Schwichtenberg *et al.*, 2016; Fawcett and Innan, 2016) and animal species (Suh *et al.*, 2017; Naville *et al.*, 2019). For SaliS-II, existence across all species analyzed here (Figure 3a) together with similar activity profiles might indicate propagation in the last common ancestor at least 65 mya. However, this contradicts the proposed high ‘turnover’ of SINE copies (Lenoir *et al.*, 2005; Baucom *et al.*, 2009; Kögler *et al.*, 2017). Instead, a genome-wide TE activity across species may be caused by common environmental influences, e.g. activation of SINEs by temperature changes. Other stress conditions, e.g. defense response to pathogens in some species populations may increase the retrotransposition

rate and promote diversification into subfamilies (Grandbastien *et al.*, 1997; Bui and Grandbastien, 2012; Negi *et al.*, 2016). Moreover, the genomic context and the chromatin status may affect the activity.

In contrast to the relatively homogeneous SaliS-II population, SaliS-IV represents an example for an extremely diverse SINE family containing species-specific SINE variants and subfamilies with highly variable activity patterns (Figure 3b, d). The four families SaliS-I, SaliS-II, SaliS-III.1 and SaliS-IV.1 are widely distributed and most likely evolved at the same time during chromosome rearrangements after the salicoid duplication. However, differentiation to species-specific SINE populations does not necessarily increase over time, exemplified by the contrasting examples of SaliS-II and SaliS-IV (Figure 3b, d). Presumably, SINE reactivation after incomplete lineage sorting is mainly responsible for the species-specific occurrence of SINE families (reviewed in Ray *et al.*, 2006), which is a frequently observed phenomenon for SINEs (Walters-Conte *et al.*, 2014; Fawcett and Innan, 2016; Jordan *et al.*, 2018).

SINE integration into genes or their regulatory sequences without harmful effects to the host have the potential to preserve a SINE copy over long evolutionary periods. For example, the wide-spread Au SINE family, in many species associated with genic regions (Ben-David *et al.*, 2013; Schwichtenberg *et al.*, 2016; Seibt *et al.*, 2016; Fawcett and Innan, 2016; Keidar *et al.*, 2018), shows high sequence conservation for at least ~ 300 million years (Magallón *et al.*, 2013), as it is present in both angiosperms and gymnosperms (Fawcett and Innan, 2016). This is in line with studies showing a preferred integration of SINEs into gene-rich regions, in particular introns (Tsuchimoto *et al.*, 2008; Baucom *et al.*, 2009; Seibt *et al.*, 2016). SINE copies settled in genic regions survive in the long-term as shown in the Solanaceae, where approximately ten percent of all annotated genes harbor at least one SINE (Seibt *et al.*, 2016).

Weakly conserved promoter motifs and relatively short poly(A) tails are sufficient for SaliS amplification

Only a few nucleotides of the SaliS promoter box motifs are conserved within the tRNA-derived promoter of the SINE 5' region (Figure 7). Thus, it is likely that novel SINE families have emerged by reshuffling (Kögler *et al.*, 2017) than by *de novo* assembly of a tRNA gene and random genomic regions with poly(A) stretches. The SaliS box A motif is more degenerated than the box B motif (two vs. five nucleotides present in all promoter motifs analyzed, Figure 7a). Consistent with other tRNA-derived plant SINE families (Wenke *et al.*, 2011), this indicates that the box B motif may be crucial for the binding of the RNA polymerase III complex (Kramerov and Vassetzky, 2005). As the weakly conserved promoter motifs are a search query for the *SINE-Finder*-based identification, some SINE families with strongly deviating box A and box B nucleotides might have escaped detection.

We found that the majority of copies within a SINE population has a relatively short poly(A) tail indicating retrotranspositional inactivity (Roy-Engel *et al.*, 2002; Odom *et al.*, 2004). With the exception of two SaliS-I copies in *P. euphratica* with tail lengths of 30 bp and 32 bp, respectively, the poly(A) tail of SaliS families ranges between 8 bp and 21 bp. Increased tail length averages (15 bp, 17 bp, and 21 bp, Table S6) might indicate recent activity as they were observed for evolutionarily young SINE families such as SaliS-IX in *S. purpurea*, SaliS-V and SaliS-IV.2 in *P. trichocarpa*, and SaliS-IV.1 in *P. euphratica*. However, the poly(A) tail lengths are not related to the diversity of SINE copies in contrast to the correlation between SINE similarity and TSD lengths (Figure 6), which was reported for Poaceae SINE families (Kögler *et al.*, 2017) and might be associated with the function of the poly(A) during retrotransposition.

The poly(A) tail mostly represents the structure shared between SINEs and LINEs (Boeke, 1997; Roy-Engel, 2012), but may also be extended upstream to homology of 3' ends of SINEs and LINEs (Okada and Hamada, 1997; Baucom *et al.*, 2009; Wenke *et al.*, 2011). The 3' poly(A) tail serves as a recognition signal for the reverse transcription by an autonomous LINE partner. Hence, it is an inherent part of a SINE and presumably not a polyadenylation product (Boeke, 1997; Dewannieux *et al.*, 2003; Borodulina *et al.*, 2016). It mediates the binding of SINE transcripts to specific proteins (e.g. poly(A) binding protein), which in turn are responsible for binding the RT of stringent LINEs

(reviewed in Okada *et al.*, 1997). Consequently, the poly(A) tail length is linked to retrotransposition efficiency (Roy-Engel *et al.*, 2002; Dewannieux and Heidmann, 2005) and, thus, affects SINE activity (Odom *et al.*, 2004). For the human Alu family, disease-causing copies have 3' tails of 40 adenine residues or more (Roy-Engel *et al.*, 2002), and it was shown that SINE activity can be rescued by tail elongation (Hagan *et al.*, 2003; Wagstaff *et al.*, 2012).

However, long poly(A) stretches are extremely unstable and shrink rapidly in size, if they are not stabilized by interruptions through single nucleotide changes (Roy-Engel *et al.*, 2002; Odom *et al.*, 2004). Thus, recently inserted SINE copies may not be inherently retrotransposition-competent (Hagan *et al.*, 2003; Deininger *et al.*, 2011).

Mutations of the 3' tail have the potential to extend the SINE length

The most striking feature of Salicaceae SINE families is the variability of their 3'ends upstream of the poly(A) tail (Figure 9), while the 5' starts are typically conserved across copies within a SINE family and sometimes even between families (this study, Figure 8, Figure S2; Schwichtenberg *et al.*, 2016; Kögler *et al.*, 2017).

We found a relationship between the type of 3' end and the age of the respective copies (Figure 9c), indicating that different 3' ends most likely emerged at different time points. These findings may be explained by different scenarios concerning the active SINE copy (putative source loci) (Cordaux *et al.*, 2004; Price *et al.*, 2004). Either a single active copy may have changed over time or a new active copy, more efficient in retrotransposition, is responsible for the altered 3' end (Britten *et al.*, 1988; Deininger and Slagel, 1988; Deininger *et al.*, 1992). Also, a few active SINEs might exist in a genome (Matera *et al.*, 1990), simultaneously producing copies corresponding to the variety of 3' ends.

In order to interpret the emergence of multiple 3' ends, we inspected tail structures and developed an evolutionary model for the enlargement of the SINE 3' region: The terminal conserved triplet of SaliS-I (GTT) is generally followed by two or three adenines (Figure 9c), presumably originating from the ancient SaliS-I poly(A) tail (Figure 11).

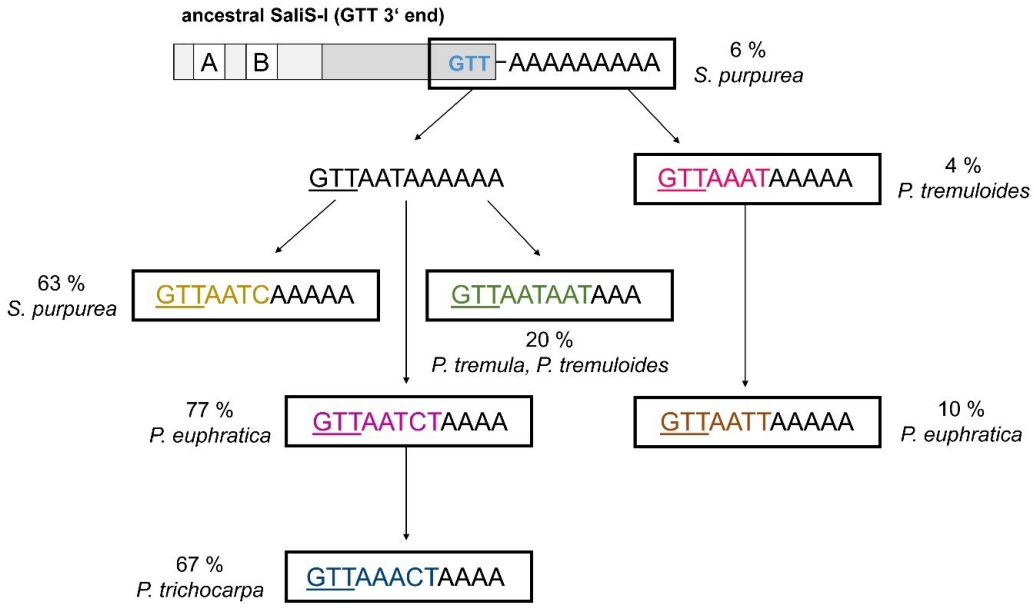


Figure 11. Development of SaliS-I 3' end variants. An ancient SaliS-I copy with the GTT 3' end gained poly(A) tail mutations at different positions, leading to the novel 3' ends AAT and AAAT. Only the 3' end motif GTTAAAT has been fixed in the SaliS-I population (fixation indicated by frames). Further 3' ends emerged with the ongoing accumulation and fixation of mutations in the 3' tail. The broadest distribution of each 3' end within the Salicaceae species analyzed is indicated with the respective frequency.

According to this scenario, GTT represents the ancestral character state of the 3' end prior to the poly(A) tail of an active SaliS-I copy. During evolution, the poly(A) tail was subject to mutations, for example the third and fourth adenine of the tail sequence. These adenine-thymine-transversions became fixed by following amplifications and led to an elongation of the SINE by three or four nucleotides derived from the 3' tail, which were then part of the SINE (GTT-AAT and GTT-AAAT, respectively, Figure 11). The 3' tail mutations might either be introduced to genomic copies or during reverse transcription, as reverse transcriptases generally lack the proofreading ability (reviewed in Hu and Hughes, 2012).

Fixation of altered 3' tail nucleotides is a result of the target primed reverse transcription. The poly(A) tail of the SINE transcript anneals to a thymine stretch at the target site, exposed after the first strand cleavage. The new SINE copy is presumably synthesized by a LINE reverse transcriptase provided in *trans* (Luan *et al.*, 1993; Ostertag and Kazazian, 2001).

Exemplified for SaliS-I's 3' end motif GTT-AAACT (Figure 11, blue), the first three adenines of the original poly(A) tail and the mutated fourth and fifth nucleotide (cytosine and thymine) became part of the SINE full-length.

These mutations may have been acquired stepwise by a master SINE copy responsible for the majority of SINE copies in the respective genome. However, it cannot be excluded that each 3' end variant originated from its own founder SINE copy as discussed above. The chronology of the mutations is not traceable, but we can assume that SaliS-I gained four (e.g. GTT-AAAT and GTT-AATC) to six (e.g. GTT-AATAAT) nucleotides, depending on the nucleotide exchange position (Figure 11).

Knowledge of SINEs as a major class of repetitive DNA sequences is crucial and constitutes an important resource for renewable energy crop genomics. Although SINEs are largely ubiquituous in all plant species investigated so far, they have only been poorly analyzed in tree species such as poplar and willow. The Salicaceae SINE landscape is formed by 20 (sub)families which diverged over evolutionary time scales. However, SaliS family abundance and sequence diversity still largely follow evolutionary key periods such as the salicoid genome duplication and the poplar-willow separation. The evolution of SINE families is promoted by:

(1) Lineage-specific differentiation of SINE families and subfamilies depending on the activity of individual diversified copies (including reactivation of ancient SINE families based on a preserved copy in genome regions of low mutation rate),

(2) Reshuffling of sequence segments between SINE families and subfamilies by nested SINE integrations or recombination events,

(3) SINE family 3' end diversification as a result of fixed poly(A) tail mutations generating subpopulations of variable 3' ends differing in sequence and length.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Argus, G.W. (2010) Salix. In Flora of North America Editorial Committee, ed. *Flora of North America North of Mexico. Vol. 7: Magnoliophyta: Salicaceae to Brassicaceae*. Oxford University Press, New York, pp. 23–162

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., SanMiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.*, **5**, e1000732.

Ben-David, S., Yaakov, B. and Kashkush, K. (2013) Genome-wide analysis of short interspersed nuclear elements SINEs revealed high sequence conservation, gene association and retrotranspositional activity in wheat. *Plant J.*, **76**, 201–210.

Blackburn, K.B. and Harrison, J.W. (1924) A preliminary account of the chromosomes and chromosome behavior in the Salicaceae. *Ann.Bot.*, **38**, 361–378.

Boeke, J. (1997) LINEs and Alus - the poly A connection. *Nat. Genet.*, **16**, 6–7.

Borodulina, O.R., Golubchikova, J.S., Ustyantsev, I.G. and Kramerov, D.A. (2016) Polyadenylation of RNA transcribed from mammalian SINEs by RNA polymerase III: Complex requirements for nucleotide sequences. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1859**, 355–365.

Boucher, L.D., Manchester, S.R. and Judd, W.S. (2003) An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. *Am. J. Bot.*, **90**, 1389–1399.

Britten, R.J., Baron, W.F., Stout, D.B. and Davidson, E.H. (1988) Sources and evolution of human Alu repeated sequences. *Evolution (N. Y).*, **85**, 4770–4774.

Bui, Q.T. and Grandbastien, M.-A. (2012) LTR Retrotransposons as controlling elements of genome response to stress? In M.-A. Grandbastien and J. M. Casacuberta, eds. *Plant transposable elements.* Springer Berlin Heidelberg, pp. 273–296.

Buzdin, A.A. (2004) Retroelements and formation of chimeric retrogenes. *Cell. Mol. Life Sci.*, **61**, 2046–2059.

Churakov, G., Smit, A.F.A., Brosius, J. and Schmitz, J. (2005) A novel abundant family of retroposed elements (DAS-SINEs) in the nine-banded armadillo (*Dasypus novemcinctus*). *Mol. Biol. Evol.*, **22**, 886–893.

Collinson, M.E. (1992) The early fossil history of Salicaceae: a brief review. *Proc. R. Soc. Edinburgh. Sect. B. Biol. Sci.*, **98**, 155–167.

Cordaux, R., Hedges, D.J. and Batzer, M.A. (2004) Retrotransposition of Alu elements: how many sources? *Trends Genet.*, **20**, 464–467.

Dai, X., Hu, Q., Cai, Q., et al. (2014) The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res.*, **24**, 1274–1277.

Deininger, P., Lander, E., Linton, L., et al. (2011) Alu elements: know the SINEs. *Genome Biol.*, **12**, 236.

Deininger, P.L. and Batzer, M.A. (1995) SINE master genes and population biology. In R. Maraia, ed. *The impact of short, interspersed elements (SINEs) on the host genome.* Georgetown, Texas: Landes, R G, pp. 43–60.

Deininger, P.L., Batzer, M.A., Hutchison, C.A. and Edgell, M.H. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.*, **8**, 307–311.

Deininger, P.L. and Slagel, V.K. (1988) Recently amplified Alu family members share a common parental Alu sequence. *Mol. Cell. Biol.*, **8**, 4566–4569.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: Origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.*, **12**, 1075–1079.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.*, **35**, 41–48.

Dewannieux, M. and Heidmann, T. (2005) LINEs, SINEs and processed pseudogenes: Parasitic strategies for genome modeling. *Cytogenet. Genome Res.*, **110**, 35–48.

Dewannieux, Marie and Heidmann, T. (2005) Role of poly(A) tail length in Alu retrotransposition. *Genomics*, **86**, 378–381**.**

Dickmann, D.I. and Kuzovkina, J. (2014) Poplars and willows of the world, with emphasis on silvicuturally important species. In J. G. Isebrands and J. Richardson, eds. *Poplars and willows.* Rome, Italy: FAO.

Du, J., Tian, Z., Bowen, N.J., Schmutz, J., Shoemaker, R.C. and Ma, J. (2010) Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR swapping in soybean. *Plant Cell*, **22**, 48–61.

Eckenwalder, J.E. (1996) Systematics and evolution of *Populus*. In R. F. Stettler, H. D. Bradshaw Jr, P. E. Heilman, and T. M. Hinckley, eds. *Biology of* Populus *and its Implications for Management and Conservation*. Ottawa, ON, Canada: NRC Research Press, pp. 7–32.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edwards, K., Johnstone, C. and Thompson, C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.*, **19**, 1349.

Fawcett, J.A. and Innan, H. (2016) High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. *Mol. Biol. Evol.*, **33**, 2593–2604.

Feschotte, C., Fourrier, N., Desmons, I., Mouches, C. and Mouchès, C. (2001) Birth of a retroposon: The twin SINE family from the vector mosquito *Culex pipiens* may have originated from a dimeric tRNA precursor. *Mol. Biol. Evol.*, **18**, 74–84.

Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: Where genetics meets genomics. *Nat. Rev. Genet.*, **3**, 329–341.

Gao, C., Xiao, M., Ren, X., Hayward, A., Yin, J., Wu, L., Fu, D. and Li, J. (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics*, **100**, 222–230.

Girgis, H.Z. (2015) Red: an intelligent, rapid, accurate tool for detecting repeats *de-novo* on the genomic scale. *BMC Bioinformatics*, **16**, 227.

Goodwin, D.C. and Lee, S. (1993) Microwave miniprep of total genomic DNA from fungi, plants, protists and animals for PCR. *Biotechniques*, **15**, 438–444.

Grandbastien, M. a, Lucas, H., Morel, J.B., Mhiri, C., Vernhettes, S. and Casacuberta, J.M. (1997) The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. *Genetica*, **100**, 241–252.

Hagan, C.R., Sheffield, R.F. and Rudin, C.M. (2003) Human Alu element retrotransposition induced by genotoxic stress. *Nat. Genet.*, **35**, 219–220.

Heslop-Harrison, J. (1991) The molecular cytogenetics of plants. *J. Cell Sci.*, **100**, 15–22.

Hou, J., Ye, N., Dong, Z., Lu, M., Li, L. and Yin, T. (2016) Major chromosomal rearrangements distinguish willow and poplar after the ancestral "salicoid" genome duplication. *Genome Biol. Evol.*, **8**, 1868–1875.

Hu, W.-S. and Hughes, S.H. (2012) HIV-1 reverse transcription. *Cold Spring Harb. Perspect. Med.*, **2**, a006882.

Johnson, L.J. and Brookfield, J.F.Y. (2006) A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Mol. Biol. Evol.*, **23**, 235–239.

Jordan, V.E., Walker, J.A., Beckstrom, T.O., et al. (2018) A computational reconstruction of *Papio* phylogeny using Alu insertion polymorphisms. *Mob. DNA*, **9**, 13.

Jühling, F., Mörl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Pütz, J. (2009) tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.

Jurka J (2010) SINE elements from black cottonwood. *Repbase Reports* **10**, 238.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V and Jurka, M. V (2005) Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. *Cytogenet. Genome Res.*, **110**, 117–123.

Kearse, M., Moir, R., Wilson, A., et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Keidar, D., Doron, C. and Kashkush, K. (2018) Genome-wide analysis of a recently active retrotransposon, Au SINE, in wheat: content, distribution within subgenomes and chromosomes, and gene associations. *Plant Cell Rep.*, **37**, 193–208.

Kögler, A., Schmidt, T. and Wenke, T. (2017) Evolutionary modes of emergence of short interspersed nuclear element (SINE) families in grasses. *Plant J.*, **92**, 676–695.

Kramerov, D.A. and Vassetzky, N.S. (2005) Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, **247**, 165–221.

Lauron-Moreau, A., Pitre, F.E., Argus, G.W., Labrecque, M. and Brouillet, L. (2015) Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS One*, **10**, e0138963.

Lawrence, M., Gentleman, R. and Carey, V. (2009) rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, 1–10.

Lenoir, A., Pélissier, T., Bousquet-Antonelli, C. and Deragon, J.M. (2005) Comparative evolution history of SINEs in *Arabidopsis thaliana* and *Brassica oleracea*: Evidence for a high rate of SINE loss. *Cytogenet. Genome Res.*, **110**, 441–447.

Levy, A., Schwartz, S. and Ast, G. (2009) Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. *Nucleic Acids Res.*, **38**, 1515–1530.

Liu, X., Wang, Z., Wang, D. and Zhang, J. (2016) Phylogeny of *Populus*-*Salix* (Salicaceae) and their relative genera using molecular datasets. *Biochem. Syst. Ecol.*, **68**, 210–215.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.

Ma, J., SanMiguel, P., Lai, J., Messing, J. and Bennetzen, J.L. (2005) DNA rearrangement in orthologous *orp* regions of the maize, rice and sorghum genomes. *Genetics*, **170**, 1209–1220.

Magallón, S., Hilu, K.H.W. and Quandt, D. (2013) Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.*, **100**, 556–573.

Manchester, S.R., Judd, W.S. and Handley, B. (2006) Foliage and fruits of early poplars (Salicaceae: *Populus*) from the Eocene of Utah, Colorado, and Wyoming. *Int. J. Plant Sci.*, **167**, 897–908.

Marco, A. and Marín, I. (2008) How Athila retrotransposons survive in the *Arabidopsis* genome.

BMC Genomics, **9**, 219.

Matera, A.G., Hellmann, U., Hintz, M.F. and Schmid, C.W. (1990) Recently transposed Alu repeats result from multiple source genes. *Nucleic Acids Res.*, **18**, 6019–6023.

Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N. and Chourrout, D. (2019) Massive changes of genome size driven by expansions of non-autonomous transposable elements. *Curr. Biol.*, **29**, 1161-1168.

Negi, P., Rai, A.N. and Suprasanna, P. (2016) Moving through the stressed genome: Emerging regulatory roles for transposons in plant stress response. *Front. Plant Sci.*, **7**, 1448.

Nishihara, H., Terai, Y. and Okada, N. (2002) Characterization of novel Alu- and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. *Mol. Biol. Evol.*, **19**, 1964–1972.

Odom, G.L., Robichaux, J.L. and Deininger, P.L. (2004) Predicting mammalian SINE subfamily activity from A-tail length. *Mol. Biol. Evol.*, **21**, 2140–2148.

Ohshima, K. and Okada, N. (2005) SINEs and LINEs: Symbionts of eukaryotic genomes with a common tail. *Cytogenet. Genome Res.*, **110**, 475–490.

Okada, N. and Hamada, M. (1997) The 3′ ends of tRNA-derived SINEs originated from the 3’ ends of LINEs: A new example from the bovine genome. *J. Mol. Evol.*, **44**, 52–56.

Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. (1997) SINEs and LINEs share common 3’ sequences: A review. *Gene*, **205**, 229–243.

Oliver, K.R., McComb, J.A. and Greene, W.K. (2013) Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.*, **5**, 1886–1901.

Ostertag, E.M. and Kazazian H.H., J. (2001) Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.*, **11**, 2059–2065.

Paesold, S., Borchardt, D., Schmidt, T. and Dechyeva, D. (2012) A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. *Plant J.*, **72**, 600–611.

Price, A.L., Eskin, E., Pevzner, P.A., Price, A.L., Eskin, E. and Pevzner, P.A. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.*, **14**, 2245–2252.

R Core Team (2017) R: A language and environment for statistical computing. Vienna, Australia: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Ragauskas, A.J., Williams, C.K., Davison, B.H., et al. (2006) The path forward for biofuels. *Science*, **311**, 484–489.

Ray, D.A., Xing, J., Salem, A.-H. and Batzer, M.A. (2006) SINEs of a nearly perfect character. *Syst. Biol.*, **55**, 928–935.

Roy-Engel, A.M. (2012) A tale of an A-tail: The lifeline of a SINE. *Mob. Genet. Elements*, **2**, 282–286.

Roy-Engel, A.M., Salem, A.H., Oyeniran, O.O., Deininger, L., Hedges, D.J., Kilroy, G.E., Batzer, M.A. and Deininger, P.L. (2002) Active Alu element "A-tails": Size does matter. *Genome Res.*, **12**, 1333–1344.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., et al. (1996) Nested retrotransposons in the intergenic regions of the maize Genome. *Science*, **274**, 765–768.

Sannigrahi, P., Ragauskas, A.J. and Tuskan, G.A. (2010) Poplar as a feedstock for biofuels : A review of compositional characteristics. *Biofuels, Bioprod., Biorefin.*, **4**, 209–226.

Schwichtenberg, K., Wenke, T., Zakrzewski, F., Seibt, K.M., Minoche, A., Dohm, J.C., Weisshaar, B., Himmelbauer, H. and Schmidt, T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229–244.

Seibt, K.M., Wenke, T., Muders, K., Truberg, B. and Schmidt, T. (2016) Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. *Plant J.*, **86**, 268–285.

Senerchia, N., Felber, F. and Parisod, C. (2014) Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. *New Phytol.*, **202**, 975–985.

Smyshlyaev, G., Voigt, F., Blinov, A., Barabas, O. and Novikova, O. (2013) Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. *Proc. Natl. Acad. Sci.*, **110**, 20140–20145.

Suh, A., Bachg, S., Donnellan, S., Joseph, L., Brosius, J., Kriegs, J.O. and Schmitz, J. (2017) *De-novo* emergence of SINE retroposons during the early evolution of passerine birds. *Mob. DNA*, **8**, 21.

Sundell, D., Mannapperuma, C., Netotea, S., et al. (2015) The plant genome iintegrative explorer resource: PlantGenIE.org. *New Phytol.*, **208**, 1149–1156.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.

Tsuchimoto, S., Hirao, Y., Ohtsubo, E. and Ohtsubo, H. (2008) New SINE families from rice, OsSN, with poly(A) at the 3' ends. *Genes Genet. Syst.*, **83**, 227–236.

Tuskan, G.A., DiFazio, S. and Jansson, S. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Ullu, E. and Tschudi, C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312,** 171–172.

Verbylaite, R., Beišys, P., Rimas, V. and Kuusiene, S. (2010) Comparison of ten DNA extraction protocols from wood of European aspen (*Populus tremula* L.). *Balt. For.*, **16**, 35–42.

Vicient, C.M. and Casacuberta, J.M. (2017) Impact of transposable elements on polyploid plant genomes. *Ann. Bot.*, **120**, 195–207.

Wagstaff, B.J., Hedges, D.J., Derbes, R.S., Campos Sanchez, R., Chiaromonte, F., Makova, K.D. and Roy-Engel, A.M. (2012) Rescuing Alu: Recovery of new inserts shows LINE-1 preserves Alu activity through A-Tail expansion. *PLoS Genet.*, **8**, e1002842.

Walters-Conte, K.B., Johnson, D.L.E., Johnson, W.E., O'Brien, S.J. and Pecon-Slattery, J. (2014) The dynamic proliferation of CanSINEs mirrors the complex evolution of feliforms. *BMC Evol. Biol.*, **14**, 137.

Wang, Z., Du, S., Dayanandan, S., Wang, D., Zeng, Y. and Zhang, J. (2014) Phylogeny reconstruction and hybrid analysis of populus (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. *PLoS One*, **9**, e103645.

Weber, B. and Schmidt, T. (2009) Nested Ty3-gypsy retrotransposons of a single *Beta procumbens* centromere contain a putative chromodomain. *Chromosom. Res.*, **17**, 379–396.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. (2016) Evolution of plant genome architecture. *Genome Biol.*, **17**, 1–14.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Wickham, H. (2016) ggplot2: Elegant graphics for data analysis (2nd Edition). Springer New York.

Wollrab, C., Heitkam, T., Holtgräwe, D., Weisshaar, B., Minoche, A.E., Dohm, J.C., Himmelbauer, H. and Schmidt, T. (2012) Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome. *Plant J.*, **72**, 636–651.

Wu, J., Gu, Y.Q., Hu, Y., You, F.M., Dandekar, A.M., Leslie, C.A., Aradhya, M., Dvorak, J. and Luo, M.C. (2012) Characterizing the walnut genome through analyses of BAC end sequences. *Plant Mol. Biol.*, **78**, 95–107.

Yadav, V.P., Mandal, P.K., Bhattacharya, A. and Bhattacharya, S. (2012) Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*. *Nat. Commun.*, **3**, 854.

Yagi, E., Akita, T. and Kawahara, T. (2011) A novel Au SINE sequence found in a gymnosperm. *Genes Genet. Syst.*, **86**, 19–25.

Chapter 3

Genotyping based on SINEs – Application of the Inter-SINE Amplified Polymorphism (ISAP) Marker System in Angiosperm and Gymnosperm Tree Species

3.1 Localization of the native East Asian origin of the Pillnitz camellia

Introduction

The Pillnitz Camellia is one of the oldest *C. japonica* trees of Europe: it was planted in 1801 at the park of Pillnitz castle and enjoys great popularity due to its early spring flowerage (Jäger, 1995). It is presumed that the first *C. japonica* specimen reached the Court of Dresden between 1770 and 1790 (Haikal, 2010) depending on conflicting theories of its origin (Haikal, 2008 and 2010). Native to East Asia, natural habitats extend from China, Taiwan, and Southern Korea to Japan (Ullmann, 2004; Mondal, 2011).

A common aspect of the two main theories is that the distribution of camellias throughout Europe most likely commenced in the United Kingdom (UK). The most famous theory is the 'Thunberg legend', reporting that the Swedish naturalist Carl Peter Thunberg brought four plants from an expedition to Japan (1775 to 1776) and donated one specimen to the Royal Botanic Gardens (Kew, UK) (Haikal, 2008). If so, these four plants might originate from the Gotō Islands, famous for large natural camellia populations and located on the main trading route between Europe and Japan of the 18th century (Dutch East India Company). However, the Royal Botanic Gardens registered a visit of Carl Thunberg in 1779 (Kümmel, 1981), while the first *C. japonica* at Kew was documented in 1789 (Aiton, 1789).

The theory of the Chinese origin points to the province Yunnan, famous for over 1,500 years of traditional cultivation of the related *C. reticulata*, which were often grown on rootstocks (Savige, 1991; Short, 2005b; Mondal, 2011; Xin *et al.*, 2015). In 1739, the collection of rare plant species of Robert James Lord Petre (Thorndon Hall, Essex, UK) was complemented by two camellias of unknown origin (Haikal, 2010). There are several indications that the plants may originate from China (Savige, 1985; Short, 2005a; Short, 2005b). The historical painting 'The peacock pheasant of China'

shows the red-flowering specimen (Edwards, 1747) and the comment ‘The flower here figured by way of decoration is called the Chinese Rose.’ (Short, 2005b). As it rather shows *C. reticulata* flowers (Figure 1), it might have been a Yunnan camellia grafted on a *C. japonica* rootstock (Short, 2005b; M. Riedel, personal communication). It is further assumed that after Lord Petre’s death in 1742 only the more robust rootstock of the camellia survived, which was subsequently propagated (Short, 2005b; Taylor, 2014; M. Riedel, personal communication).

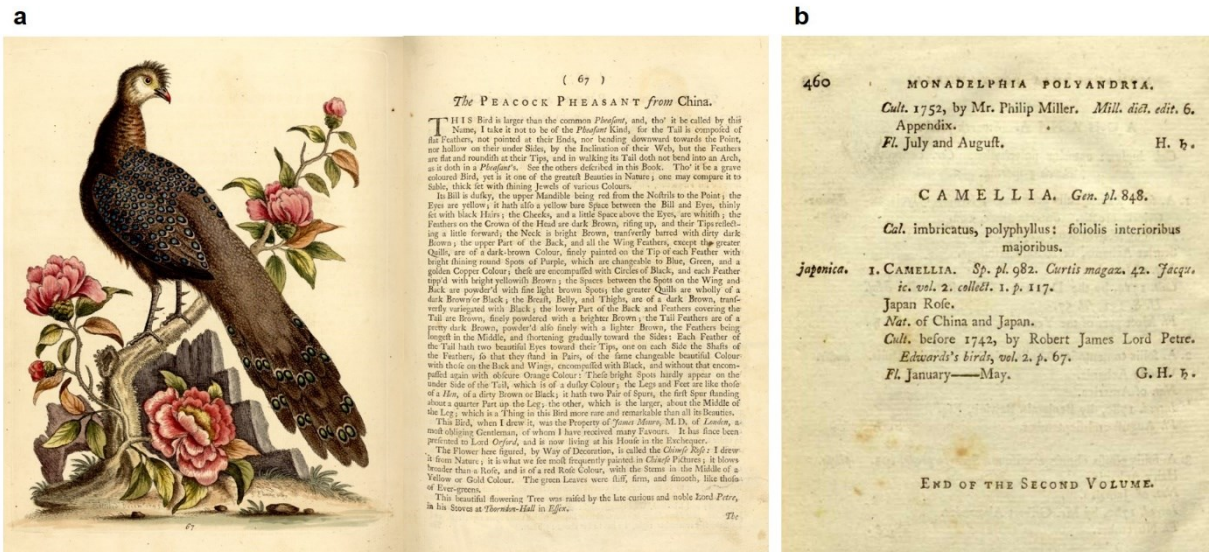


Figure 1. Historical documents supporting the theory of the Chinese origin of the Pillnitz Camellia. (a) A painting of the probably first living camellia on European ground was published 1747 in George Edward’s *A natural history of birds*. (b) The first evidence of a *C. japonica* specimen at the Royal Botanic Garden (Kew) is dated back to 1789. The author William Aiton noted ‘cultivated before 1742, by Robert James Lord Petre’.

In order to uncover the geographical origin of the Pillnitz Camellia, an ISAP analysis of numerous candidate *C. japonica* genotypes was performed. The collection contains other old European camellias, established cultivars, accessions from the potential regions of origin, and genotypes of similar phenotype. Based on the assumption that natural Japanese *C. japonica* populations might be distinguishable from those of China, a tendency is expected pointing to one of the two mostly discussed origins.

The ’Schlösser, Burgen und Gärten Sachsen gemeinnützige GmbH’ (Dresden, Germany) provided funding for the elucidation of the geographical origin of the Pillnitz camellia using the SINE-based marker system ISAP. This work was realized at the chair of Plant Cell and Molecular Biology and the research group Molecular and Organismic Diversity at the chair of Botany of the Dresden University

of Technology (Dresden, Germany) in collaboration with Matthias Riedel, curator of the camellia collection at the Landschloss Pirna-Zuschendorf (Pirna, Germany).

Experimental procedures

Plant material and DNA preparation

Genomic DNA of *Camellia* genotypes (Table 1) was extracted from lyophilized leaf material using different commercial kits. Mainly, the kit 'NucleoSpin Plant II' (Macherey-Nagel, Düren, Germany) was used with exception of the *C. japonica* leaf material originating from the Gotō Islands (Japan). The DNA of these samples was isolated with the 'DNeasy Plant Maxi Kit' (Qiagen, Valencia, US).

Each DNA extraction was followed by ethanol precipitation using 1/10 volume of 3 M sodium acetate and 2.5 volumes of ethanol. The samples were incubated overnight at - 20 ºC. Centrifugation was carried out at 14,000 rpm and 4 °C for 30 minutes. After carefully decanting the supernatant, the DNA pellet was rinsed twice with 75 % ethanol (diluted with distilled water), followed by centrifugation for 5 minutes each. The DNA pellet was air-dried at 50 °C using the ThermoMixer (Eppendorf, Hamburg, Germany). The DNA was dissolved in distilled water and stored at - 20ºC until usage.

DNA quality control

The purity of the genomic DNA samples was estimated by measurement of the absorbance ratio 260 nm / 280 nm (A260/A280) using the NanoDrop™ spectrophotometer (Implen, München, Germany). DNA solutions with the A260/A280 quotient of 1.8 to 2.0 were considered to be pure.

The DNA integrity was verified by electrophoretical separation of 3 µl of genomic DNA in solution with 7 µl of distilled water and 2 µl of 6x loading dye (Thermo Scientific, Waltham, USA). Intact DNA was recognizable as a single band of high molecular weight.

Table 1. Camellia genotypes used for ISAP analysis. The classification was taken from Chang and Bartholomew (1984). Unknown origin is indicated by N/A.

No.	Subgenus	Section	Species	Cultivar / name	Origin	Source
1	*Camellia*	*Camellia*	*Camellia japonica*	PKS-1TU	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
2	*Camellia*	*Camellia*	*Camellia japonica*	PKS-2TU	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
3	*Camellia*	*Camellia*	*Camellia japonica*	PKS-3TU	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
4	*Camellia*	*Camellia*	*Camellia japonica*	PKS-4TU	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
5	*Camellia*	*Camellia*	*Camellia japonica*	PKS-5TU	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
6	*Camellia*	*Camellia*	*Camellia japonica*	Pillnitz	Pillnitz, Germany	Landschloss Pirna-Zuschendorf [a]
7	*Camellia*	*Camellia*	*Camellia japonica*	Campo Bello	Campo Bello, Portugal	Landschloss Pirna-Zuschendorf [a]
8	*Camellia*	*Camellia*	*Camellia japonica*	Caserta	Caserta, Italy	Landschloss Pirna-Zuschendorf [a]
9	*Camellia*	*Camellia*	*Camellia japonica*	Greifswald	Greifswald, Germany	Landschloss Pirna-Zuschendorf [a]
10	*Camellia*	*Camellia*	*Camellia japonica*	Ashiya	Ashiya, Japan	Landschloss Pirna-Zuschendorf [a]
11	*Camellia*	*Camellia*	*Camellia japonica*	Chidan	Chidan, China	Landschloss Pirna-Zuschendorf [a]
12	*Camellia*	*Camellia*	*Camellia japonica*	Althaeiflora	N/A	Landschloss Pirna-Zuschendorf [a]
13	*Camellia*	*Camellia*	*Camellia japonica*	Mathotiana Alba	Belgium	Landschloss Pirna-Zuschendorf [a]
14	*Camellia*	*Camellia*	*Camellia japonica*	Tricolor	N/A	Landschloss Pirna-Zuschendorf [a]
15	*Camellia*	*Camellia*	*Camellia japonica*	Herme	Japan	Landschloss Pirna-Zuschendorf [a]
16	*Camellia*	*Camellia*	*Camellia japonica*	Alba plena	N/A	Landschloss Pirna-Zuschendorf [a]
17	*Camellia*	*Camellia*	*Camellia japonica*	Magnoliaeflora	N/A	Landschloss Pirna-Zuschendorf [a]
18	*Camellia*	*Camellia*	*Camellia japonica*	Toki-Hime	Gotō Islands, Japan	Goto Camellia Forest Park [b]
19	*Camellia*	*Camellia*	*Camellia japonica*	Osako No. 1	Gotō Islands, Japan	Goto Camellia Forest Park [b]
20	*Camellia*	*Camellia*	*Camellia japonica*	Onidake	Gotō Islands, Japan	Goto Camellia Forest Park [b]
21	*Camellia*	*Camellia*	*Camellia japonica*	Hoso-Goryô	Gotō Islands, Japan	Goto Camellia Forest Park [b]
22	*Camellia*	*Camellia*	*Camellia japonica*	Virgin Maria	Gotō Islands, Japan	Goto Camellia Forest Park [b]
23	*Camellia*	*Camellia*	*Camellia japonica*	Kunming	Kunming, China	Botanical Gardens of Kunming University [c]
24	*Camellia*	*Camellia*	*Camellia japonica ssp. rusticana*	Rusticana	N/A	Landschloss Pirna-Zuschendorf [a]
25	*Camellia*	*Camellia/ Oleifera*	*C. japonica* × *C. sasanqua*	Sayohime	N/A	Landschloss Pirna-Zuschendorf [a]
26	*Camellia/ Metacamellia*	*Camellia/ Theopsis*	*C. japonica* × *C. lutchuensis*	Sweet Emily Kate	N/A	Landschloss Pirna-Zuschendorf [a]
27	*Camellia/ Metacamellia*	*Camellia/ Theopsis*	*C. japonica* × *C. lutchuensis*	Scentuous	N/A	Landschloss Pirna-Zuschendorf [a]
28	*Camellia*	*Oleifera*	*Camellia sasanqua*	Floribunda	N/A	Landschloss Pirna-Zuschendorf [a]
29	*Camellia*	*Paracamellia*	*Camellia grijsii*	Villa Orsi	N/A	Landschloss Pirna-Zuschendorf [a]
30	*Thea*	*Thea*	*Camellia sinensis*	O. Kuntze	N/A	Landschloss Pirna-Zuschendorf [a]

[a] (Pillnitz, Germany) / [b] (Gotō Islands, Japan) / [c] N 25°8'18.197" E 102° 44'39.719" (Kunming, China)

The DNA digestibility was tested with the FastDigest™ restriction endonuclease *Bsu*RI (Thermo Scientific, Waltham, USA). The reaction mixture was prepared as follows and incubated at 37 °C for 15 minutes.

Reaction mix:

Genomic DNA (0.5 - 1 µg)	5.0 µl	
Distilled water	11.0 µl	
10x FastDigest™ Green Buffer	2.0 µl	
FastDigest™ endonuclease *Bsu*RI*	2.0 µl	* concentration (U/µl) not provided
Total volume	20.0 µl	

ISAP PCR and agarose gel electrophoresis

For the development of ISAP markers basically any SINE family of a genome can be used. However, to achieve a high degree of selectivity for the discrimination between *C. japonica* genotypes, SINE families with high abundance and high similarity are an important prerequisite. As this decision had to be made at an early stage of the SINE identification progress, ISAP primers were derived from the SINE cluster with the highest sequence count (Chapter 2.1, Table 3). These SINE cluster correspond to the SINE families TheaS-I to TheaS-IV (Chapter 2.1, Table 4). For each of these four SINE families two outward-facing primers were derived to enable the amplification of the flanking genomic regions between adjacent SINE copies by polymerase chain reaction (PCR) (Table 2). As the primers were developed based on SINEs of the *C. japonica* genome, they were designated CjS (*Camellia japonica* SINE).

Table 2. ISAP primer. For standard PCR the 20mer primers were used as listed. For the ISAP PCR the SINE-derived primers were elongated by a 5' GC-rich extension (5' - CTGACGGGCCTAACGGAGCG - 3') resulting in 40mer primers.

SINE family	***forward* Primer**		***reverse* Primer**	
	name	sequence (5' - 3' orientation)	name	sequence (reverse complement)
TheaS-I	CjS-I_*for*	GAGGATAGGGAGGATTTTCC	CjS-I_*rev*	GGGTGCCTGTTAGCCGTTCC
TheaS-II	CjS-II_*for*	TACTCAATCTTTCCCCTCCC	CjS-II_*rev*	AATGCACAAAGTGGTTGCCC
TheaS-III	CjS-III_*for*	CAGGGATTAGTCGAGGTGCG	CjS-III_*rev*	AGCTCTGTATGGACTGGCCC
TheaS-IV	CjS-IV_*for*	GATGACACCTCAGAGCATCC	CjS-IV_*rev*	ACCACACGCCACAGACAAGC

The ISAP primers consist of a 20mer SINE-derived region, ensuring the SINE specificity of the ISAP bands, and a 20mer GC-rich region of arbitrary sequence, equal for all primers, that enables the application of the two-step ISAP PCR (Table 2). Furthermore, to avoid their binding to tRNA genes based on the highly conserved 11 bp motifs of the tRNA-derived promotor (box A – TGGCnnAGTGG and box B - GGTTCGAnnCC; Galli *et al.*, 1981), the ISAP primers were preferentially derived from the SINE 3' region. All primers were developed using the browser-based tool *OligoAnalyzer* (IDT, Coralville, US) and obtained from Eurofins Genomics (Ebersberg, Germany).

Each DNA sample was tested in dilution series with the *C. japonica*-specific ISAP primers in a standard PCR to determine the optimal concentration for an informative banding pattern. The ingredients of the PCR mixture were adopted from Wenke *et al.* (2015).

PCR ingredients:

Genomic DNA (~ 20 ng/µl)	1.0 µl
Distilled water	9.7 µl
10× DreamTaq™ Green Buffer	2.0 µl
dNTPs (2 mM)	2.0 µl
BSA (bovine serum albumin) (2 mg/ml)	2.0 µl
Betaine (50 mM)	1.0 µl
ISAP primer 1 (10 µM)	1.0 µl
ISAP primer 2 (10 µM)	1.0 µl
DreamTaq™ DNA polymerase (5 U/µl)	0.3 µl
Total volume	20.0 µl

The DreamTaq™ DNA polymerase (Thermo Scientific, Waltham, USA) provided the sharpest bands compared to other frequently used DNA polymerases, e.g. GoTaq® DNA polymerase (Promega, Madison, USA). Thermal cycling was performed on the Mastercycler epgradient S (Eppendorf, Hamburg, Germany). The individual steps of the standard PCR and the ISAP program (Wenke *et al.*, 2015) were as follows:

Standard PCR program:

94 °C	5 min		initial denaturation
94 °C	20 s	30 ×	denaturation
z °C	30 s		annealing
72 °C	2 min		extension
72 °C	5 min		final extension
4 °C	∞		storage

ISAP PCR program:

93 °C	5 min		initial denaturation
93 °C	20 s	3 ×	denaturation
z °C	30 s		annealing
72 °C	2 min		extension
93 °C	20 s	27 ×	denaturation
72 °C	140 s		annealing/extension
72 °C	5 min		final extension
4 °C	∞		storage

Opposed to the standard PCR, the ISAP PCR has a dual composition: the first three cycles also consist of three steps, including the primer annealing to the DNA template ('z' is usually 50 - 56 °C). For the following 27 cycles, annealing and elongation are fused to a single step at 72 °C. Based on the GC-rich 5' extension and the higher temperature, the ISAP primers only bind to already synthesized amplicons of the preceding cycles. As a result, the intensity of small-sized bands decreases and more large-sized bands can be amplified.

For the separation of PCR products ethidium bromide (0.05 µl/ml gel) stained agarose gels were run in 1 × TAE at 60 V – 80 V, using the Sub-Cell ® GT Agarose Gel Electrophoresis Systems and power supplies from BioRad (Berkeley, USA). For testing DNA integrity and digestibility 1.2 % agarose gels were prepared, while ISAP products were separated using 2 % of agarose (Seakem® LE, Lonza, Rockland, US). The complete reaction volume of the PCR (20 µl) was loaded onto the gel and the size standard (2 µl) 'GeneRuler™ 100 bp Plus DNA Ladder' (Thermo Scientific, Waltham, USA) was added.

50× TAE buffer:

2 M Tris base
2 M glacial acetic acid
50 mM EDTA dissolved in distilled water
pH 8.5

diluted to 1× TAE with distilled water

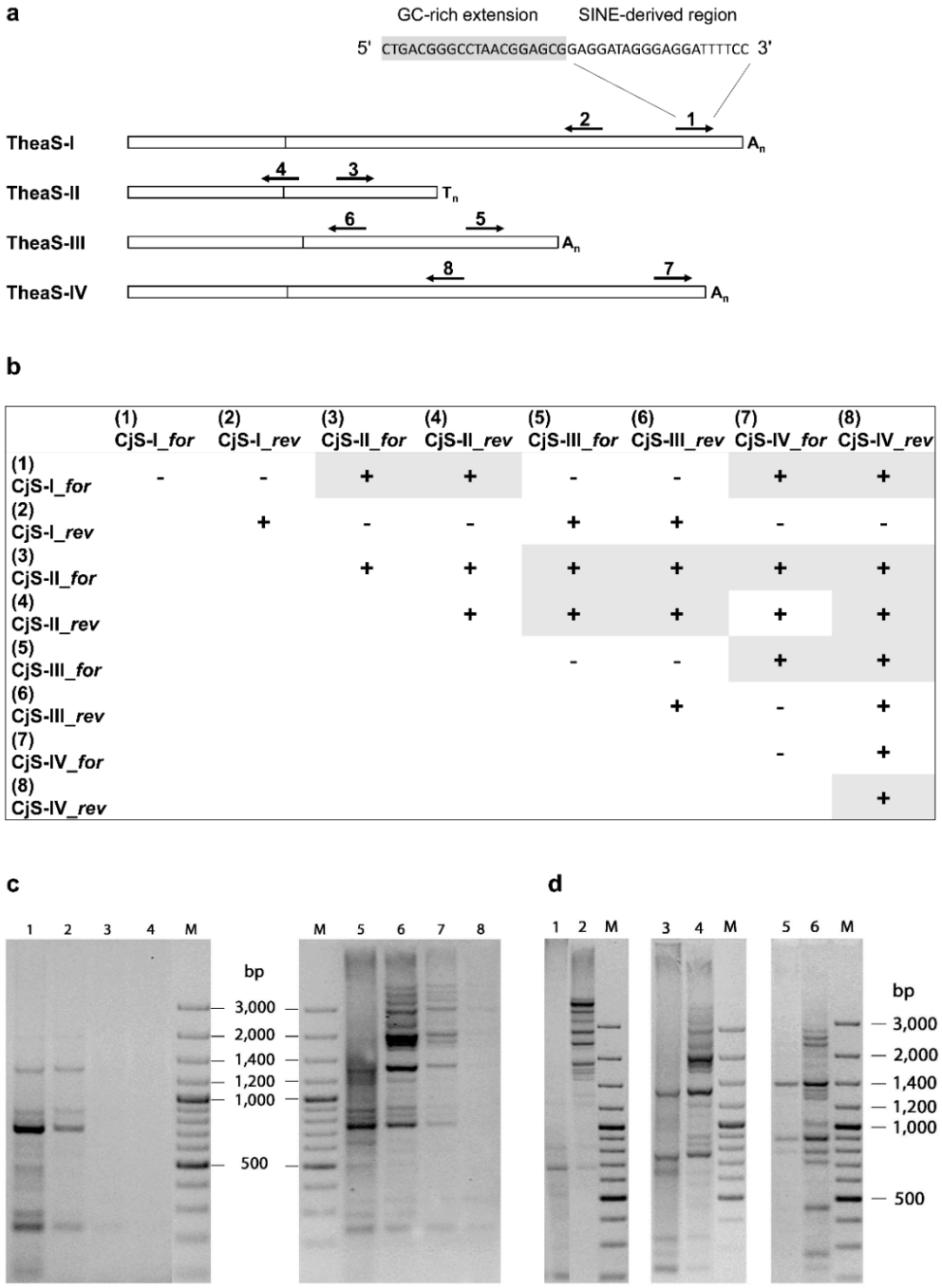
ISAP analysis

The gel images were captured with the Gel Doc™ 2000 Gel Documentation System (Bio-Rad, USA) and the ISAP banding patterns were analyzed with *GelCompar II* (Applied Maths NV, Belgium). The ISAP analysis includes the normalization of banding patterns according to the size standard and the automated band classification. The resulting band size classes contain the information "band present", "band absent" or "uncertain" for weak bands. The cluster analysis of combined ISAP data was performed using the unweighted pair-grouping with arithmetic mean (UPGMA) based on Dice similarity coefficients. Dendrograms were constructed using 1000 bootstrap repetitions.

Results

Establishment of the ISAP marker system for *Camellia japonica*

Eight *C. japonica* ISAP primers were tested with genomic DNA of the Pillnitz camellia as reference genotype (Figure 2a; Table 2). These CjS (*Camellia japonica* SINE) primers can be applied individually or combined in pairs, resulting in 36 possible combinations (Figure 2b). The resulting PCR amplicons form a specific banding pattern ('fingerprint'), as shown exemplarily for the primer combination CjS-I_*for* / CjS-II_*rev* (Figure 2c, 1-4). The number of bands was increased by optimization of annealing temperatures and application of the specialized ISAP PCR (Figure 2c, 5 - 8). The quality and quantity of the DNA significantly influences the reproducibility and comparability of the banding patterns and were tested accordingly (Figure 2d). Each DNA extraction was complemented by an additional ethanol precipitation step to obtain high-purity genomic DNA (Figure 2d, 1 - 2). The number of bands is reduced, if the template DNA is added insufficiently or in large amounts. Best results were achieved using 16 - 45 ng of DNA for a PCR assay (Figure 2d, 3 - 6).



	(1) CjS-I_*for*	(2) CjS-I_*rev*	(3) CjS-II_*for*	(4) CjS-II_*rev*	(5) CjS-III_*for*	(6) CjS-III_*rev*	(7) CjS-IV_*for*	(8) CjS-IV_*rev*
(1) CjS-I_*for*	-	-	+	+	-	-	+	+
(2) CjS-I_*rev*		+	-	-	+	+	-	-
(3) CjS-II_*for*			+	+	+	+	+	+
(4) CjS-II_*rev*				+	+	+	+	+
(5) CjS-III_*for*					-	-	+	+
(6) CjS-III_*rev*						+	-	+
(7) CjS-IV_*for*							-	+
(8) CjS-IV_*rev*								+

Figure 2. Design of ISAP primers and optimization of the banding patterns. (a) The position of ISAP primers is indicated on the respective TheaS families. A vertical line separates the SINE 5' region from the 3' region. A 20mer GC-rich extension is added to the SINE-derived ISAP primer, preferentially derived from the SINE 3' region. (b) The primer combinations producing bands (+) were tested under ISAP PCR conditions to select those with more than four bands (grey shading) to detect marker candidates. (c) Initially, the primers were tested in a standard gradient PCR (annealing temperatures: 1 – 50 °C, 2 – 53 °C, 3 – 56 °C, 4 – 59 °C, M – size marker 'GeneRuler™ 100 bp Plus DNA Ladder') and subsequently applied under ISAP PCR conditions (annealing temperatures: 5 – 50 °C, 6 – 52 °C, 7 – 54 °C, 8 – 56 °C). (d) The influence of DNA quality (1-2) and concentration (3-6) on ISAP banding patterns is shown: ISAP PCR with CjS-I_*for* / CjS-II_*rev* and genomic 'Toki-Hime' DNA before ethanol precipitation (lane 1, 40 ng) and after (lane 2, 30 ng); ISAP PCR with CjS-I_*for* / CjS-II_*rev* and genomic 'Pillnitz' DNA (lane 3 – 90 ng; lane 4 – 45 ng); ISAP PCR with CjS-II_*rev* / CjS-IV_*rev* and genomic 'Mathotiana Alba' DNA (lane 5 – 8 ng; lane 6 – 16 ng).

After adaptation of the reaction conditions and evaluation of the individual results, 14 primer sets showing a sufficient number of bands with sufficient resolution on the agarose gel were selected as marker candidates (Figure 3).

Compared to the results of the standard PCR, the ISAP PCR has the potential to increase the number of bands. Mostly, additional bands occur, which consist of larger amplicons (Figure 3, e.g. 1/3, 1/7, 3/7, 4/5, 4/6, 4/8, 5/7). Other examples show depletion of small-sized bands for the benefit of large-sized bands (1/4, 1/8, 8/8). Three banding patterns (3/5, 3/8, 5/8) could not be clearly improved. The primer combination 3/6 shows identical banding patterns in both PCR assays. The primer combination 1/8 exhibits bands of the size range 300 bp - 1000 bp, which are detectable with standard PCR, but too faint for a clear evaluation using the ISAP PCR. The whole range of inter-SINE PCR bands would only be accessible by combination of both patterns.

The most efficient ISAP primers for informative banding patterns are CjS-I_*for* (1), CjS-II_*for* (3), and CjS-IV_*rev* (8), contributing to four primer combinations each (Figure 2b). The usage of single ISAP primers like CjS-IV_*rev* (8) is less efficient for *C. japonica* genotyping, as the distance between neighboring TheaS copies is most likely too large to generate an adequate ISAP banding pattern (Figure 3, 8/8).

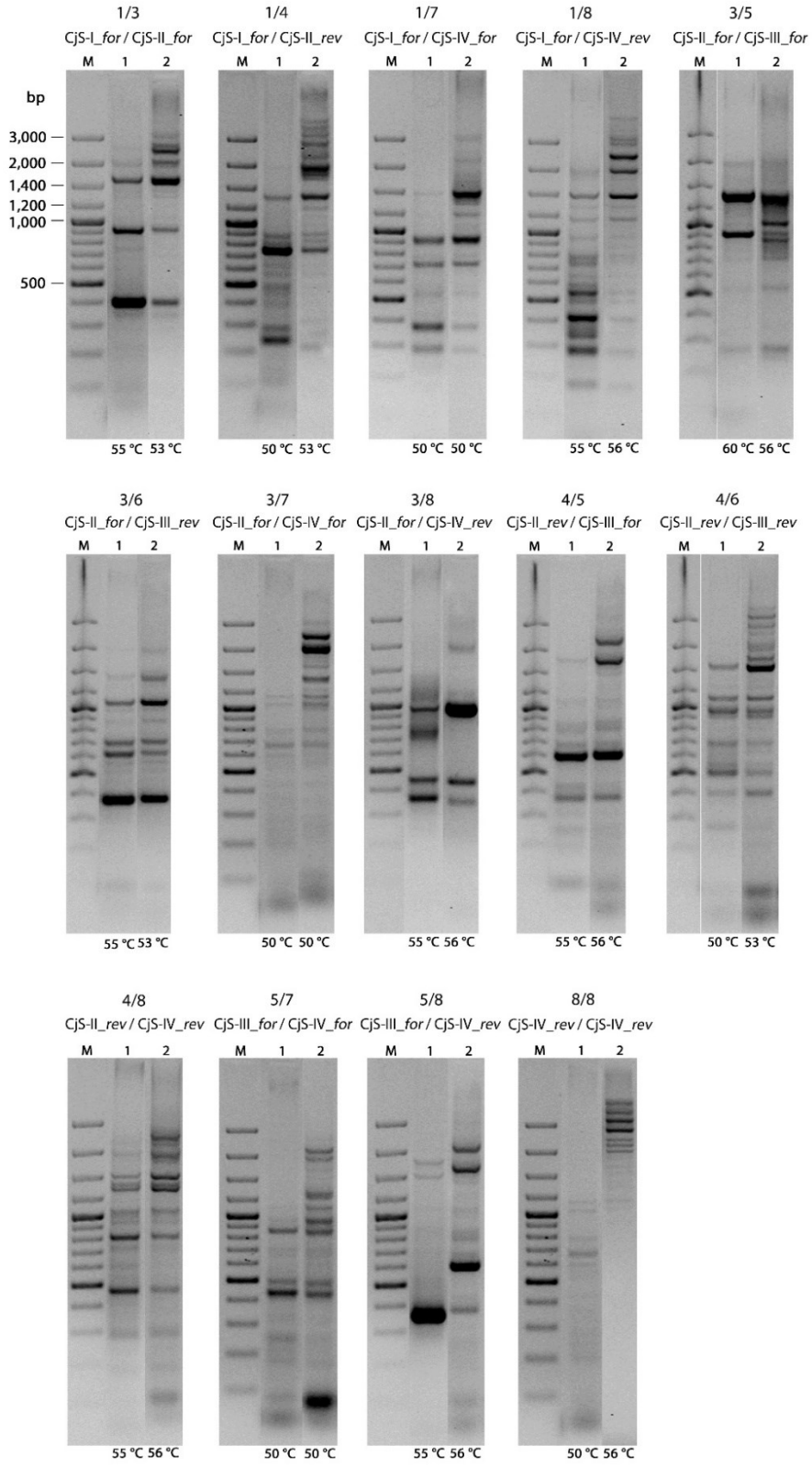


Figure 3. ISAP marker candidates derived from the Pillnitz camellia (*C. japonica*). The annealing temperatures are indicated below each lane for the standard PCR (1) and the ISAP PCR (2), each. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

For the development of ISAP markers, highly polymorphic banding patterns have to be identified. Thus, the marker candidates were tested with seven *C. japonica* genotypes (Pillnitz, Caserta, Campobello, Ashiya, Chidan, Althaeiflora, and the subspecies *rusticana*). Based on the combined ISAP data of four informative primer combinations, a UPGMA cluster analysis was performed to visualize the genetic diversity of the *Camellia* genotypes investigated (Figure 4a). Three *Camellia* species (*C. sasanqua*, *C. grijsii*, and *C. sinensis*) were included to measure the percentage similarities to more distantly related genotypes.

Four *C. japonica* genotypes show identical fingerprints for all four ISAP primer combinations tested (Figure 4b, 1 - 4), indicating that these plants originate from the same *C. japonica* specimen and were propagated vegetatively. Three of them are old European camellia trees, cultivated for at least 200 years (Vela *et al.*, 2009), and the fourth is a presumed snow camellia (*C. japonica* subsp. *rusticana*), which was included due to its similar habitus and flower morphology compared to the Pillnitz plant (Figure 4a, black areas).

The genotypes Ashiya and Chidan were examined as *Camellia* representatives of Japanese and Chinese origin, respectively. The plant referred to the genotype ‘Ashiya’ originates from a camellia forest above the city Ashiya between Osaka and Kyoto. It was imported by the curator of the Zuschendorf camellia collection (M. Riedel) in 1989. The genotype ‘Chidan’ is supposed to origninate from China and the *C. japonica* cultivar Althaeiflora is one of the oldest European varieties, dated back to 1824 (M. Riedel, personal communication). These three gennotypes are similar to the first group of genetically identical plants with 67 % - 83 % identity (Figure 4a, dark blue areas). The fingerprints of the three species *C. sasanqua*, *C. grijsii*, and *C. sinensis* differ from the *C. japonica* genotypes investigated, showing mostly 50 % - 66 % identity to this group (Figure 4a, blue areas).

The application of additional ISAP primer combinations could substantiate the similarities reflecting genetic relationships. However, evaluable fingerprints including polymorphic bands could only be achieved for the four primer combinations indicated in Figure 4b.

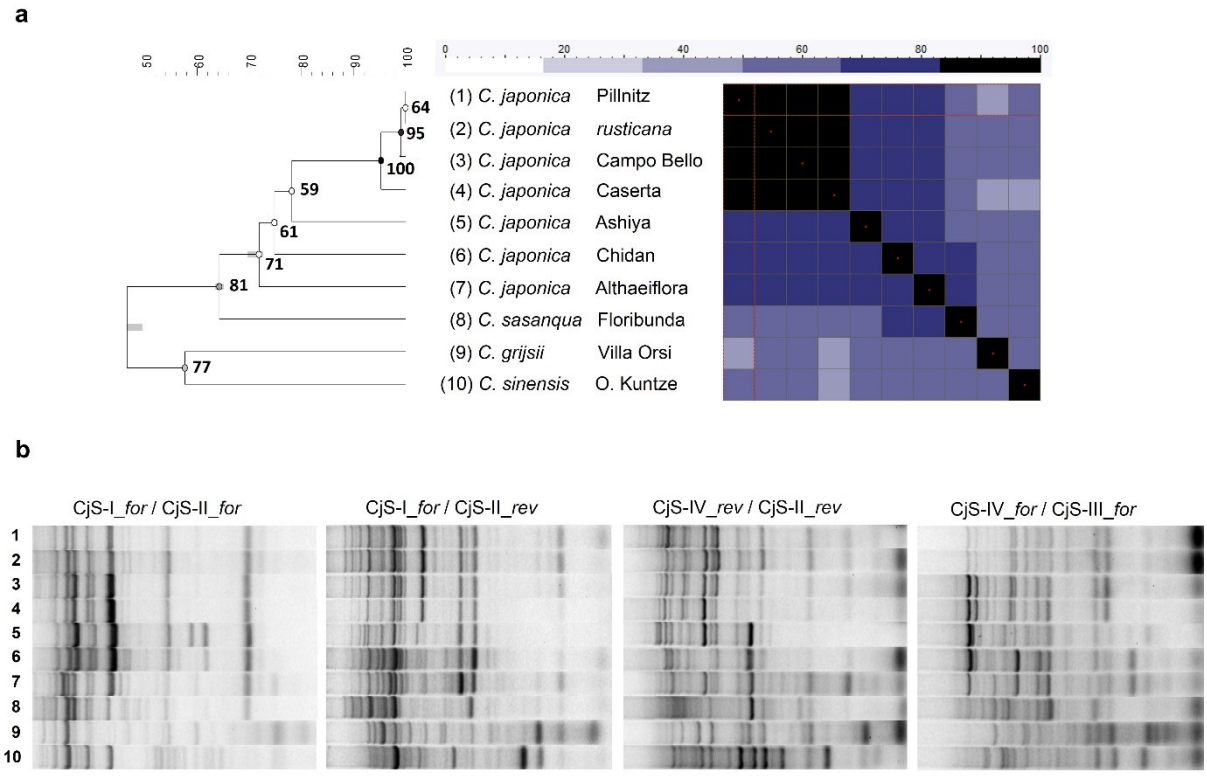
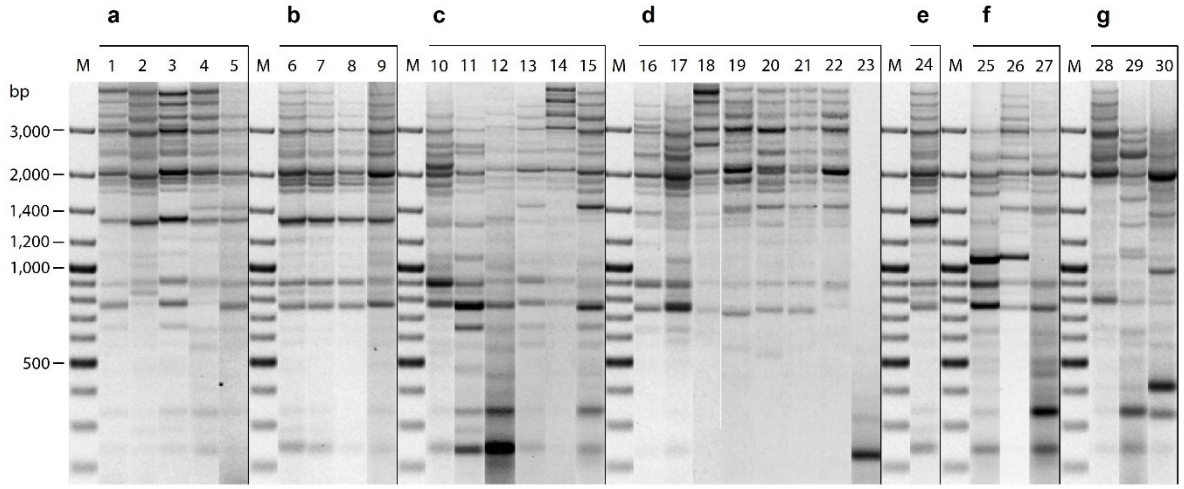


Figure 4. Genetic variability among *Camellia* genotypes indicates a common origin of Europe's oldest *C. japonica* trees. (a) The percentage similarities between *Camellia* genotypes were determined by UPGMA cluster analysis with the dice similarity coefficient for band matching and are presented as color-coded matrix. The dendrogram branch quality was calculated with 1000 bootstrap simulations. (b) The cluster analysis is based on fingerprints of four different primer combinations.

Genetic diversity of *Camellia* genotypes

A collection of 30 *Camellia* accessions was analyzed with the primer combination CjS-I_*for* / CjS-II_*rev* to analyze the genetic diversity of differently related groups of *Camellia* genotypes reflected by their ISAP banding patterns (Figure 5). The collection contains descendants of the Pillnitz camellia resulting from self-pollination (Figure 5a), historical European camellias (*C. japonica*) originating from the 18th century (Figure 5b), popular ornamental *C. japonica* cultivars (Figure 5c), *C. japonica* samples of potential geographical origins (Figure 5d), the *C. japonica* subspecies *rusticana* (Figure 5e), interspecific *C. japonica* hybrids (Figure 5f), and three other *Camellia* species (Figure 5g).



(a) Seed-propagated plants of the Pillnitz camellia tree	
1 *Camellia japonica*	PKS-TU1
2 *Camellia japonica*	PKS-TU2
3 *Camellia japonica*	PKS-TU3
4 *Camellia japonica*	PKS-TU4
5 *Camellia japonica*	PKS-TU5
(b) Old European *Camellia japonica* trees	
6 *Camellia japonica*	Pillnitz (Germany)
7 *Camellia japonica*	Campo Bello (Portugal)
8 *Camellia japonica*	Caserta (Italy)
9 *Camellia japonica*	Greifswald (Germany)
(c) Popular *Camellia japonica* cultivars	
10 *Camellia japonica*	Althaeiflora
11 *Camellia japonica*	Mathotiana Alba
12 *Camellia japonica*	Tricolor
13 *Camellia japonica*	Herme
14 *Camellia japonica*	Alba Plena
15 *Camellia japonica*	Magnoliaeflora

(d) *Camellia japonica* plants of native Asian origin	
16 *Camellia japonica*	Ashiya (Japan)
17 *Camellia japonica*	Chidan (China)
18 *Camellia japonica*	Toki-Hime (Japan)
19 *Camellia japonica*	Osako No.1 (Japan)
20 *Camellia japonica*	Onidake (Japan)
21 *Camellia japonica*	Hoso Goryô (Japan)
22 *Camellia japonica*	Virgin Maria (Japan)
23 *Camellia japonica*	Kunming (China)
(e) *Camellia japonica* subspecies	
24 *Camellia japonica*	*rusticana*
(f) Inter-species *Camellia japonica* hybrids	
25 *C. japonica* × *C. lutchuensis*	Scentuous
26 *C. japonica* × *C. lutchuensis*	Sweet Emily Kate
27 *C. japonica* × *C. sasanqua*	Sayohime
(g) *Camellia* species	
28 *Camellia sasanqua*	Floribunda
29 *Camellia grijsii*	Villa Orsi
30 *Camellia sinensis*	O. Kuntze

Figure 5. The ISAP profiles of 30 *Camellia* genotypes show variability in accordance with genetic relationships. The analysis is based on primer combination CjS-I_*for* / CjS-II_*rev*. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

The seedlings of the Pillnitz Camellia resulting from self-pollination show only minor genetic variability as expected. Compared to the pattern of the Pillnitz camellia (Figure 5b, 6), new bands were detected, e.g. at ~ 1,400 bp in PKS-TU4 and PKS-TU5 or at ~ 3,100 bp in PKS-TU1 and PKS-TU3 and the absence of others was observed, e.g. at ~ 1,800 bp in PKS-TU5 or at 850 bp in PKS-TU1 and PKS-TU3 (Figure 5a).

The group of old European camellia trees show identical ISAP banding patterns (Figure 5b). Therefore, it is likely that these plants are genetically identical and were presumably propagated by cuttings. The variability among Zuschendorf cultivars is hardly to interpret due to DNA quality problems (Figure 5c) and the genetic heterogeneity of this historical cultivar collection, consisting of various cultivars imported from numerous geographical locations, often followed by subsequent breeding. However, fundamental differences between their ISAP profiles were not observed and

regarding the size range of 2,000 bp to 3,000 bp they are more similar to each other than to the Pillnitz-type ISAP pattern.

The analysis of native Asian *C. japonica* samples available in this study is shown in Figure 5d. The genetic diversity of five different *C. japonica* specimens from the Gotō Islands of Japan (Toki-Hime, Osako No.1, Onidake, Hoso Goryô, Virgin Maria), resembling the phenotype of the Pillnitz camellia (M. Riedel, personal communication)(Riedel, 2016)(Riedel, 2016)(Riedel, 2016) was examined. These samples originate from old trees (250 - 400 years) growing at the Gotō Camellia Forest Park, famous for a wide range of camellia species and cultivars. The ISAP patterns of the Gotō *C. japonica* trees (Figure 5d, 18 - 22) show only low variability, comparable to the Pillnitz camellia seedlings or the Zuschendorf cultivars, and might therefore exhibit a higher degree of relationship. However, they clearly differ from the Pillnitz-type ISAP profile in the size range of 1,200 bp - 1,400 bp (Figure 5b, 6 - 9; Figure 5d, 18 - 22).

The ISAP banding pattern of the genotype Ashiya (Figure 5d, 16), also of Japanese origin, differs significantly from the Gotō samples in the range above 3,000 bp. Remarkably, the ~ 3,000 bp band, present in most *C. japonica* ISAP patterns, is absent in Ashiya, which shows a band below and above the 3,000 bp marker band instead. The genotype Chidan (Figure 5d, 17), as a *C. japonica* representative of Chinese origin, differs less from the Japanese *C. japonica* genotypes, the Gotō samples and the Ashiya genotype.

Nevertheless, the ISAP profiles of the Gotō trees are distinguishable from the Pillnitz-type ISAP pattern, which makes it unlikely that the Pillnitz camellia is a vegetatively propagated descendant of a Gotō specimen investigated here.

To test the hypothetical Chinese origin of the Pillnitz camellia, leaf material of a *C. japonica* specimen growing at the Botanical Gardens of Kunming University (Kunming, province Yunnan, China) was analyzed. An extraction of intact DNA failed, presumably due to the storage of the leaves on silica gel during transport, instead of the required immediate freeze-drying to avoid DNA degradation. Hence, it was not possible to generate an evaluable ISAP profile (Figure 5d, 23).

The snow camellia is a subspecies of *C. japonica* (*C. japonica* subsp. *rusticana* (Honda) Kitamura), which is morphologically adapted to heavy snow fall and naturally mainly occurs at altitudes of 350 to

1,000 meters above sea-level (Kume and Tanaka, 1996). However, the specimen designated ‘snow camellia’ shares the same ISAP profile with the four old European camellias investigated (Figure 5b, e) and therefore might have been mislabeled.

The 22 *C. japonica* ISAP profiles demonstrate the presence of species-specific bands, e.g. the 3,000 bp band and the three strong bands above (Figure 5a-e). However, DNA quality fluctuations influence the respective banding patterns and complicate the comparison: some ISAP profiles show distinct bands mainly between 1,000 bp and 3,000 bp (Figure 5c, 10 - 12), others rather between 200 bp and 1,000 bp (Figure 5c, 14). The *C. japonica* cultivar Mathotiana Alba shows balanced band intensities in the size range of 200 bp to 3,000 bp (Figure 5c, 13).

The three interspecific *Camellia* hybrids, comprising *C. japonica* genome portions, but also those of the *Camellia* species *C. sasanqua* and *C. lutchuensis*, respectively, still show many typical *C. japonica* bands (Figure 5f). ISAP banding patterns are also generated in other species of the genus, exemplified by specimens of *C. sasanqua*, *C. grijsii* and *C. sinensis* (Figure 5g). Only few putative *C. japonica*-specific bands were observed in *C. grijsii* and *C. sinensis*, whose ISAP banding patterns also differ greatly from each other. The comparison of the ISAP profile of the hybrid cultivar ‘Sayohime’ (*C. japonica* × *C. sasanqua*) with those of the *C. japonica* genotypes (Figure 5a-e) and the ISAP profile of *C. sasanqua* (Figure 5g, 28) reveals many bands typical for *C. japonica*. The hybrid cultivars ‘Scentuous’ and ‘Sayohime’ exhibit highly similar ISAP banding patterns, although differing in one parental species (*C. lutchuensis* and *C. sasanqua*, respectively). Contrary, ‘Scentuous’ and ‘Sweet Emily Kate’, both *C. japonica* × *C. lutchuensis* hybrids, show greater differences in their ISAP banding patterns.

The majority of bands within the *C. japonica* ISAP profiles is species-specific and therefore less informative. Genetically identical *Camellia* accessions like vegetatively propagated specimens are easily detectable, as they show identical ISAP profiles. The analysis of the Pillnitz camellia descendants resulting from selfing, the Japanese Gotō individuals, and the Zuschendorf cultivars indicate only slightly variable ISAP banding patterns among each other.

Discussion

The identification, characterization and differentiation of *C. japonica* cultivars based on morphological traits is ineffective and needs to be complemented by molecular marker techniques. Thousands of registered cultivars (Lombard *et al.*, 2001; Couselo *et al.*, 2010; The Online Camellia Register, 2019) place high demands on the marker resolution.

In *Camellia*, efforts in marker development mainly focused on the tea plant *C. sinensis* due to its economical importance (Tripathi and Negi, 2006). The genotyping of *C. japonica* accessions, ecotypes and germplasm collections (Ueno *et al.*, 2000; Caser *et al.*, 2010; Lin *et al.*, 2013; Zhao *et al.*, 2017) are mainly based on species-specific microsatellite markers. However, high similar inner-population identities rapidly increase the number of required SSRs for differentiation (Vela *et al.*, 2013).

For the determination of the geographical origin of the Pillnitz camellia tree, the marker resolution, which means selectivity for genotype discrimination, within and among natural populations plays a major role. The ISAP marker system was established for *Camellia japonica* to evaluate the resolution between groups of differently related genotypes.

Two outward-facing primers each were derived from four of 15 TheaS families (Figure 2a), populating the *C. japonica* genome with more than 400 copies, showing at least 71 % similarity (Chapter 2.1, Table 4). Of 36 examined ISAP primer combinations (Figure 2b), 14 were considered as marker candidates (Figure 3) and four highly polymorphic banding patterns were selected for the analysis of the genetic similarity of seven *C. japonica* genotypes and three *Camellia* species (Figure 4b). The UPGMA cluster analysis thereof revealed genetic identity of three old European camellia trees and their doubtful accordance with the *C. japonica* subspecies *rusticana* (Figure 4a). The *C. japonica* genotypes of native Japanese and Chinese origin, Ashiya and Chidan, respectively, show a similar genetic distance to this group, as well as the Zuschendorf cultivar Althaeiflora (Figure 4a).

The Greifswald camellia shares a common ancestor with Europe's three oldest camellia trees

The ISAP analysis revealed that the ~ 100 year old *C. japonica* tree of the Botanical Garden Greifswald (Greifswald, Germany) (Oberdörfer, 2016; Supplementary chapter, Figure S1) and Europe's oldest three *C. japonica* trees at Pillnitz (Dresden, Germany), Caserta (Caserta, Italy), and Campo Bello (Vila Nova de Gaia, Portugal) (Savige, 1985) most likely originate from the same stock plant. Recordings of the Botanical Garden Greifswald indicate that the camellia is the descendant from an older specimen, which was imported from England in 1791 (Oberdörfer, 2016). The commercially available camellias from England of the late 18th century probably originate from the red 'Lord Petre camellia'. After Lord Petre's death in 1742, camellias were traded on a grand scale (Short, 2005b). From this point, the track of the first living camellia on European ground branches out:

(I)

Camellia specimens were commercially introduced by Lord Petre's friend and gardener Phillip Miller. The German gardener Johannes Busch, who spent some years of apprenticeship with Miller, later commenced his own business in London and adopted Miller's assortment. In 1771, he sold the camellia nursery to Conrad Loddiges, who continued and expanded the plant trading business (Haikal, 2010). Since Miller's catalogues (e.g. *Camellia Japonica flore Maximo Roseo* from 1777) were well known in Germany, Loddiges is a potential supplier of the original Greifswald camellia (M. Riedel, personal communication).

(II)

However, James Gordon, the gardener of Lord Petre, had already propagated the camellia at Petre's lifetime and offered suckers thereof in his nursery in Mile End (London) since 1942. A later partner of this nursery, Johann Andreas Graefer, was involved in the design of an English landscape garden as part of the Royal Gardens of Caserta (Italy) since 1786, commissioned by Maria Carolina of Austria, Queen of Naples and Sicily. Hence, it is likely that the Italian camellia originally came from the Gordon nursery (M. Riedel, personal communication).

As the Pillnitz camellia most likely descends from Lord Petre's red flowering camellia, indication is given for the Chinese origin (Savige, 1985; Short, 2005a; Short, 2005b). The historical painting of this camellia (Edwards, 1747) is supplemented with the designation 'Chinese Rose'. Hence, the origin of this plant might presumably be passed down orally. Furthermore, the *C. japonica* suckers, traded in England since 1742, were most likely propagated from the rootstock of the original importet plant, showing *C. reticulata*-type flowers (Short, 2005b; M. Riedel, personal communication). This grafting technique is associated with the characteristical cultivation tradition of camellias in the province Yunnan of China (Savige, 1991; Short, 2005b; Mondal, 2011; Xin *et al.*, 2015).

However, the Chinese origin of Lord Petre's camellias is highly speculative and therefore, the geographical origin of the Pillnitz camellia yet remains unknown. The Japanese origin is not excluded, although none of the analyzed old *C. japonica* individuals from the Gotō Islands is the decendant specimen.

Diagnostic size ranges within the ISAP patterns display the among-population diversity

The three oldest European camellia trees (Savige, 1985), growing at Pillnitz Castle Park (Pillnitz, Germany), Quinta de Campo Belo (Vila Nova de Gaia, Portugal), and Naples' Caserta Park (Caserta, Italy), were most likely propagated vegetatively from the same ancestor *C. japonica* specimen by cuttings or sucker. Their genetic identity is indicated by four ISAP markers (Figure 5b) and was already stated by Vela *et al.* (2009) based on 14 SSR markers. The ISAP analysis revealed that the old camellia tree at the Botanical Garden Greifswald (Greifswald, Germany) also originates from this lineage (Figure 5b).

The five descendants of the Pillnitz camellia, resulting from self-pollinated seeds show mainly four polymorphic band classes at 850 bp, 1,400 bp, 1,800 bp and 3,100 bp (Figure 5a). The genetic constitution of the Pillnitz camellia is not known. Assuming heterozygosity, these results might mainly be achieved by meiotic recombination.

The up to 400 year old *C. japonica* trees from the Japanese Gotō Islands, designated Toki-Hime, Osako No.1, Onidake, Hoso Goryô, and Virgin Maria also exhibit only slight varying ISAP banding patterns (Figure 5d, 18-22). However, they are clearly distinct from the Pillnitz-type ISAP pattern

(Figure 5b) in the ‘diagnostic size range’ of 1,200 bp to 1,400 bp. Thus, the descent from one of these old trees is unlikely. The *C. japonica* genotype Ashiya, also native to Japan and resembling the Pillnitz phenotype, exhibits two characteristic bands, above and below 3,000 bp, also clearly differing from the Pillnitz-type ISAP pattern. This might probably constitute a diagnostic region of the ISAP profile to delimit specimens of the Ashiya region.

These investigations indicate that the differentiation of geographical origins might be feasible using ISAP, moreover, as native *C. japonica* populations show high genetic variability among populations most likely due to geographic isolation (Wendel and Parks, 1985; Lin *et al.*, 2013; Nybom, 2004). In the ISAP experiments, the within-population diversity is evident from polymorphic bands distributed over the whole fingerprint area, while the among-population variability is rather displayed in specific diagnostic size ranges. Remarkably, the differentiation of Chinese and Japanese populations was demonstrated by ISSR analysis, thereby detecting a low inner-population genetic diversity (Lin *et al.*, 2013).

Cultivar differentiation of the Seidel collection

As observed for the Gotō camellia group, the Zuschendorf cultivars can be distinguished from the Pillnitz-type genotypes by a diagnostic size range, which is between 2,000 bp to 3,000 bp (Figure 5c). The samples of these cultivars were provided by Matthias Riedel, curator of Germany’s largest camellia collection at the Landschloss Pirna-Zuschendorf (Pirna, Germany), comprising more than 200 cultivars. They originate from the historical ‘Seidel camellia collection’ of Dresden, collected at the beginning 19th century by Johann Heinrich Seidel. A catalogue of Seidel’s camellia nursery from 1846 listed 540 camellia cultivars (Riedel and Riedel, 2005). He imported cultivars, such as ‘Herme’ (Japan), ‘Mathothiana’ (Belgium) and ‘Chandler’s Elegans’ (England), but also produced new varieties (Riedel and Riedel, 2005).

Thus, compared to the Gotō camellia group and the Pillnitz camellia seedlings, a higher variability within the ISAP profiles of the Zuschendorf cultivars was expected. Likely, some polymorphic bands might have not been detected due to unbalanced band intensity (Figure 5c, 12, 14) or generally faint bands of ISAP profiles resulting from insufficient DNA quality. For an increase of resolution and thus,

the significance of the results, an intended ISAP analysis based on four polymorphic primer combinations could not be performed, as the necessary optimization of each DNA sample is laborious and not feasible in the given timeframe.

However, some Zuschendorf cultivars might possess higher similarity to each other, due to the 'bud sports' phenomenon (Foster and Aranzana, 2018). Single shoots of a plant sometimes show a novel phenotype, which is stable in cuttings thereof and result from spontaneous somatic mutations in meristematic cells. As an example, in 1956 a pink flowering branch of the white flowering cultivar 'Chandler's Elegans' was introduced as cultivar 'Bernhard Lauterbach' to honor an expert in morphological cultivar identification (Riedel and Riedel, 2005). Today, the knowledge required for differentiation between the 80-100-year-old camellias of the Seidel collection is missing.

Hence, molecular techniques might facilitate cultivar differentiation of the original genetic material preserved in Zuschendorf as shown for old *C. japonica* specimens of historical gardens in Spain, Portugal, Italy, UK, Belgium and Germany (Vela *et al.*, 2013). Redundant and mislabeled accessions can be identified, as shown for the declared snow camellia, which is in fact a descendant of the Pillnitz camellia (Figure 5b, e). Presumably, the plants were confused during the hasty transport of the Seidel camellia collection from the original Seidel nursery in Dresden-Laubegast to Zuschendorf in the early 1990s to ensure the survival of the plants (Riedel and Riedel, 2005). In the course of the German reunification the Seidel company was closed.

Variable genome compositions of interspecific *Camellia* hybrids

Interspecific *Camellia* hybrids result from crosses of different *Camellia* species. Hybrid cultivars, derived from the same parent species can vary greatly in their ISAP profiles (Figure 5f, 25-26), as observed for different species (Figure 5g). This is associated with dramatically variable genome compositions in hybrid species (Langdon *et al.*, 2018). Hence, the ISAP profile of the hybrid cultivar 'Sayohime' (*C. japonica* × *C. sasanqua*), characteristic for *C. japonica* genotypes, might be explained by a major contribution of the *C. japonica* parent.

Hybrid studies might be another application area for the rapid, cost-effective ISAP method. New interspecific hybrids are still developed to obtain superior properties like all year-round bloom (Jiyin *et al.*, 2014), caused by altered gene expression patterns and transposable element mobility (Zhang *et al.*, 2018). The comparison of the parent fingerprints with those of a large progeny sample pool might reveal major genome contributions.

The geographical origin of the Pillnitz camellia remained unresolved due to the low number of comparable native Asian samples. However, Lin *et al.* (2013) demonstrated that natural *C. japonica* populations are distinguishable using molecular markers.

Microsatellite-derived markers (Ueno *et al.*, 1999; Abe *et al.*, 2006) provided insights into the genetic variability and age structure of *C. japonica* populations (Ueno *et al.*, 2000; Ueno *et al.*, 2002; Chung *et al.*, 2003) and revealed an increased among-population genetic variability (Lin *et al*., 2013), while the within-population variability depends on the geographical location as shown recently by Ryu *et al.* (2019) using a combined technique of AFLP and cpDNA regions.

This study revealed the potential to determine the *C. japonica* among-population diversity based on specific diagnostic size ranges within the ISAP banding patterns (Figure 5).

The ISAP resolution might be enhanced by including additional primer combinations. The four polymorphic primer pairs applied for genotype comparisons (Figure 4) are based on TheaS-I to TheaS-IV. The abundance of these SINE famililes ranges from 428 (TheaS-I) to 1,328 (TheaS-II) total copies, and the similarity ranges from 71 % to 81 % (Chapter 2.1, Table 5). Of the remaining *C. japonica* SINEs, the four TheaS families TheaS-VII, TheaS-VIII.1, TheaS-X, and TheaS-XIII have similar properties and might serve for primer design as well.

To clarify the origin of the Pillnitz camellia, a large sample pool of *C. japonica* plants resembling the Pillnitz phenotype from different regions of Japan and China has to be analyzed. However, the sampling of native Asian *C. japonica* specimens is difficult and requires the opportunity and the permission to collect the samples on-site. Moreover, the leaf material immediately has to be freeze-dried to avoid DNA degradation during transport. Consequently, this project requires international scientific cooperation.

References

Abe, H., Matsuki, R., Ueno, S., Nashimoto, M. and Hasegawa, M. (2006) Dispersal of *Camellia japonica* seeds by *Apodemus speciosus* revealed by maternity analysis of plants and behavioral observation of animal vectors. *Ecol. Res.*, **21**, 732.

Aiton, W. (1789) *Hortus Kewensis; or, a catalogue of the plants cultivated in the Royal Botanic Garden at Kew*, London: Printed for George Nicol, Bookseller to his Majesty, p. 460.

Caser, M., Torello Marinoni, D. and Scariot, V. (2010) Microsatellite-based genetic relationships in the genus *Camellia*: potential for improving cultivars. *Genome*, **53**, 384–399.

Chung, M.Y., Epperson, B.K. and Gi Chung, M. (2003) Genetic structure of age classes in *Camellia japonica* (Theaceae). *Evolution*, **57**, 62–73.

Couselo, J., Vela, P., Salinero, C. and Sainz, M. (2010) Characterization and differentiation of old *Camellia japonica* cultivars using simple sequence repeat (SSRs) as genetic markers. In *International Camellia Congress*. Kurume, Japan.

Edwards, G. (1747) *A natural history of birds, Volume 2*, London: Royal College of Physicians, p. 53.

Foster, T.M. and Aranzana, M.J. (2018) Attention sports fans! The far-reaching contributions of bud sport mutants to horticulture and plant biology. *Hortic. Res.*, **5**, 44.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Haikal, M. (2008) *Das Geheimnis der Kamelie* [In German], Dresden: Sandstein Verlag.

Haikal, M. (2010) *Der Kamelienwald: Die Geschichte einer deutschen Gärtnerei* [In German], Dresden: Sandstein Verlag.

Jäger, S. (1995) Das Wirken des Hofgärtners Carl Adolf Terscheck in Dresden [In German]. In *Mitteilungen des Landesvereins Sächsischer Heimatschutz e.V. - Heft 1*. Dresden: Landesverein Sächsischer Heimatschutz e.V., pp. 31–35.

Jiang, G.-L. (2013) Molecular markers and marker-assisted breeding in plants. In S. B. Andersen, ed. *Plant breeding from laboratories to fields.* Rijeka: IntechOpen.

Jiyin, G., Xinkai, L., Naisheng, Z. and Danfeng, Y. (2014) A new generation of *Camellia* hybrids. *Int Camellia J*, **46**, 49–51.

Kume, A. and Tanaka, C. (1996) Adaptation of stomatal response of *Camellia rusticana* to a heavy snowfall environment: winter drought and net photosynthesis. *Ecol. Res.*, **11**, 207–216.

Kümmel, F. (1981) The oldest camellias in the German democratic republic. *Am Camellia Yearb*, **36**, 164–175.

Langdon, Q.K., Hittinger, C.T., Peris, D. and Kyle, B. (2018) sppIDer: a species identification tool to investigate hybrid genomes with high-throughput sequencing. *Mol. Biol. Evol.*, **35**, 2835–2849.

Lin, L., Ni, S. and Li, J.-Y. (2013) Genetic diversity of *Camellia japonica* (Theaceae), a species endangered to East Asia, detected by inter-simple sequence repeat (ISSR). *Biochem. Syst. Ecol.*, **50**, 199–206.

Lombard, V., Dubreuil, P., Dilmann, C. and Baril, C. (2001) Genetic distance estimators based on molecular data for plant registration and protection: a review. In *Acta Horticulturae.* International Society for Horticultural Science (ISHS), Leuven, Belgium, pp. 55–63.

Mondal, T. (2011) *Camellia*. In C. Kole, ed. *Wild crop relatives: genomic and breeding resources, plantation and ornamental crops*. Springer Berlin Heidelberg, pp. 15–39.

Nybom, H. (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.*, **13**, 1143–1155.

Oberdörfer, E. (2016) Nordeuropas älteste Kamelie steht im Botanischen Garten [In German]. *Ostsee-Zeitung*, 1–3.

Riedel, M. [Matthias] and Riedel, M. [Marion] (2005) Saxony's Camellias. *Int Camellia J*, **37**, 69–77.

Ryu, Y., Kim, I.R., Su, M.H., Jung, J., Choi, H.-K. and Kim, C. (2019) Phylogeographical study of *Camellia japonica* inferred from AFLP and chloroplast DNA haplotype analyses. *J. Plant Biol.*, **62**, 14–26.

Savige, T. (1991) New Chinese reticulatas. *Int. Camellia J.*, **23**, 70–71.

Savige, T. (1985) The ancient camellias of Europe. *Int Camellia J*, **17**, 80–82.

Short, H. (2005a) England's first camellias. *Int Camellia J*, **37**, 51–56.

Short, H. (2005b) The truth about Lord Petre's camellias. *Int Camellia J*, **37**, 56–59.

Taylor, J.M. (2014) *Visions of loveliness: great flower breeders of the past*, Ohio University Press.

Tripathi, S.B. and Negi, M.S. (2006) Molecular markers as tools for characterization and improvement of tea germplasm. *Int. J. tea Sci.*, **5**, 29–37.

Ueno, S., Tomaru, N., Yoshimaru, H., Manabe, T. and Yamamoto, S. (2000) Genetic structure of *Camellia japonica* L. in an old-growth evergreen forest, Tsushima, *Japan. Mol. Ecol.*, **9**, 647–656.

Ueno, S., Tomaru, N., Yoshimaru, H., Manabe, T. and Yamamoto, S. (2002) Size-class differences in genetic structure and individual distribution of *Camellia japonica* L. in a Japanese old-growth evergreen forest. *Heredity*, **89**, 120.

Ueno, S., Yoshimaru, H., Tomaru, N. and Yamamoto, S. (1999) Development and characterization of microsatellite markers in *Camellia japonica* L. *Mol. Ecol.*, **8**, 335–346.

Ullmann, H.F. (2004) *Botanica: the illustrated A-Z of over 10,000 garden plants and how to cultivate them*, Könemann.

Vela, P., Couselo, J., Salinero, C., González, M. and Sainz, M. (2009) Morpho-botanic and molecular characterization of the oldest camellia trees in Europe. *Int Camellia J*, **41**, 51–57.

Vela, P., Salinero, C., Couselo, J.L., Paz, C., González-García, M. and Sainz, M.J. (2013) Characterization of *Camellia japonica* cultivars using molecular markers. *Int Camellia J*, **45**, 61–70.

Wendel, J.F. and Parks, C.R. (1985) Genetic diversity and population structure in *Camellia japonica* L. (Theaceae). *Am. J. Bot.*, **72**, 52–65.

Wenke, T., Seibt, K.M., Döbel, T., Muders, K. and Schmidt, T. (2015) Inter-SINE Amplified Polymorphism (ISAP) for rapid and robust plant genotyping. In J. Batley, ed. *Plant genotyping: methods and protocols*. New York: Springer, pp. 183–192.

Xin, T., Riek, J. de, Guo, H., Jarvis, D., Ma, L. and Long, C. (2015) Impact of traditional culture on *Camellia reticulata* in Yunnan, China. *J. Ethnobiol. Ethnomed.*, **11**, 74.

Zhang, M., Liu, X.-K., Fan, W., Yan, D.-F., Zhong, N.-S., Gao, J.-Y. and Zhang, W.-J. (2018) Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from *Camellia*. *Genome*, **61**, 477–485.

Zhao, Y., Ruan, C., Ding, G.J. and Mopper, S. (2017) Genetic relationships in a germplasm collection of *Camellia japonica* and *Camellia oleifera* using SSR analysis. *Genet Mol Res.*, **16**, doi: 10.4238/gmr16019526.

3.2 Identification of fast-growing, high yielding *Populus* genotypes for cultivation in short rotation coppice (SRC) plantations

Introduction

In order to meet the increasing demand for renewable energy sources, the fast-growing tree species poplar and willow are grown in SRC plantations (Pontailler *et al.*, 1999; van Dam *et al.*, 2007; Yemshanov and McKenney, 2008; Dillen *et al.*, 2013; Niemczyk *et al.*, 2016). Cultivated poplar clones are usually harvested after four to six years (Baum *et al.*, 2009) and the ability of the rootstocks to re-sprout enables several harvests until new plantings are required (Eppler *et al.*, 2007; Vanbeveren *et al.*, 2017). Compared to other poplar species, the European aspen (*Populus tremula* L.) is more tolerant to harsh environmental conditions, such as nutrient-poor soils and dry climate (Leibundgut, 1967; Mohrdiek, 1977; Lasch *et al.*, 2010). Especially rapid juvenile growth is reported for interspecific poplar hybrids (Heräjärvi and Junkkonen, 2006; Sixto *et al.*, 2014; Pearce *et al.*, 2018). In order to increase biomass production, high-yielding clones of suitable poplar hybrids (e.g. *P. tremula* × *P. tremuloides;* Liesebach *et al.*, 1999; Lin *et al.*, 2018) have to be identified to develop commercial cultivars. According to the *International Poplar Commission*, registered poplar cultivars are generally ‘clones’ (Dickmann and Isebrands, 2001).

For this purpose, poplar hybrids are investigated with respect to their performance in SRC plantations and depending on specific environmental and climate requirements (Lasch *et al.*, 2010; Sixto *et al.*, 2014; Liesebach, 2015). However, due to frequent hybridizations across the genus, poplars are genetically highly diverse (Floate, 2004; DiFazio *et al.*, 2011) and require high-resolution molecular markers to distinguish between complex hybrid clones.

The joint project ‘Development of retrotransposon-based molecular marker for the identification of varieties, clones and accessions as a basis for breeding, management of resources and quality control for poplar and hybrid larch’ (short title ‘TreeSINE’) aims to examine the potential of the ISAP marker system to resolve hybrid poplar accessions. The collaboration of the Dresden University of Technology, including the chair of Plant Cell and Molecular Biology (Dresden, Germany) and the group Molecular Physiology of Woody Plants (Tharandt, Germany) with the Saxony State Forestry

Service (Pirna, Germany) is funded by the program ‘Renewable raw materials’ of the German Federal Ministry of Food and Agriculture (BMEL), which is coordinated and supervised by the Agency for Renewable Resources e.V. (FNR).

The aim of this chapter is to differentiate between hybrid poplar clones cultivated on a SRC testing area of the TreeSINE partner Saxony State Forestry Service. For this purpose, a collection of different Salicaceae genotypes from *Populus,* but also *Salix,* was comparatively analyzed with *P. tremula* ISAP primers to examine the marker applicability in related species. The respective ISAP profiles were compiled to a fingerprint catalogue as a basis for storage, comparisons and evaluation of the polymorphism density and can be supplemented by commercial clones, ecotypes and collections of wild accessions.

Experimental procedures

Plant material and DNA preparation

Genomic DNA of *P. tremula* accessions analyzed in this study was obtained by the group Molecular Physiology of Woody Plants of the Dresden University of Technology (Tharandt, Germany). The accession 7590 was used as a reference for initial ISAP experiments and the accessions 10720-I, 10719-I, 10718-I, 10717-I, 10713-I, 10711-II, 10701-I, 10696-I, 10686-II, and 7589 were used for ISAP analyzes.

ISAP PCR and agarose gel electrophoresis

The ISAP experiments were carried out as described in Chapter 3.1. The ISAP primers derived from the SaliS families of *P. tremula* are listed in Table 1.

Table 1. ISAP primer. For standard PCR the 20mer primers were used as listed. For the ISAP PCR the SINE-derived primers were extended by a 5' GC-rich extension (5' - CTGACGGGCCTAACGGAGCG - 3') resulting in 40mer primers.

SINE family	***forward* Primer**		***reverse* Primer**	
	name	sequence (5' - 3' orientation)	name	sequence (reverse complement)
SaliS-I	PtS-I_*for*	AGCTGGCCCGGACACCCACG	PtS-I_*rev*	CACCACGACTAATCCCACGG
SaliS-III.1	PtS-III.1_*for*	CCTGGACCCACAAAATACGC	PtS-III.1_rev	CGGCTGTCCCAGGCTCTTAC
SaliS-IV.3	PtS-IV.3_*for*	GGTCGTTAACTTCAGGGCCC	PtS-IV.3_*rev*	CCTCTTGGTCCCAAGCTCTT
	PtS-IV.1a_*for*	CCTGTCACCCCCGCGGTGCC	PtS-IV.1_*rev*	GCACCGCGGGGGTGACAGGC
	PtS-IV.1b_*for*	ATGCTCACTGGGTTTGCAGG		

ISAP analysis

The ISAP banding patterns were analyzed and comparatively evaluated as described in Chapter 3.1. However, the DNA fingerprint software *BioNumerics* (Applied Maths NV, Belgium) was applied, which is based on *GelComparII* (Chapter 3.1), but also provides additional features like the opportunity to complement the fingerprint analysis by phenotype data. ISAP experiments comparing accessions of different poplar species and hybrid poplar clones were performed twice. The evaluation of the respective banding patterns was restricted to a region from 200 bp to 2000 bp to minimize the influence of frequently occurring weak bands above 2 kb in size.

Results

The SINE analysis in the Salicaceae (Chapter 2.3) revealed robust conditions for the establishment of the ISAP technique in the four relevant poplars for ISAP primer design (*Populus trichocarpa*, *Populus deltoides*, *Populus tremula*, and *Populus tremuloides*). The number of Salicaceae SINE (SaliS) families and subfamilies ranges between seven (*P. deltoides*) and nine (*P. tremuloides*) (Chapter 2.3, Figure 1). Four of them (SaliS-I, SaliS-II, SaliS-III.1, and SaliS-IV.1) are shared between these poplars (except SaliS-IV.1 missing in *P. deltoides*) and exhibit sufficient copy numbers for an ISAP primer design.

Due to higher relevance for economical applicability and a solid basis of interdisciplinary breeding research in Germany (www.fastwood.org; Liesebach, 2015) the European aspen (*P. tremula)* was proposed for ISAP primer design by the group Molecular Physiology of Woody Plants of the Dresden University of Technology (Tharandt, Germany).

Development of ISAP markers for *Populus tremula*

In the European aspen (*Populus tremula* L.) especially SaliS-I, SaliS-II, and SaliS-IV.1 are highly amplified with 1,174, 1,922 and 902 full-length copies, respectively (Chapter 2.3, Figure 1), and might therefore provide an adequate resource for the development of ISAP primers (Table 1). However, the high 3' end sequence similarities of the SaliS families have to be considered for the primer design (Figure 1a). The two outward-facing *P. tremula* SINE (PtS) primers derived from SaliS-I most likely can also bind to copies of SaliS-II, SaliS-IV.1, SaliS-IV.3, and SaliS-VI.1 (Figure 1a, orange region, 84 % - 96 % similarity). Despite the lower copy number, primers were derived from SaliS-I, as SaliS-II mainly consists of diverged copies (Chapter 2.3, Figure 3). The primers derived from SaliS-IV.1 might also bind to SaliS-VI.1 copies, as these SaliS families share a 57 bp central region of 97 % sequence identity (Figure 1a, blue region). Similar to abundance and similarity of the SINE families, this strongly affects the number and size distribution of bands in the banding patterns.

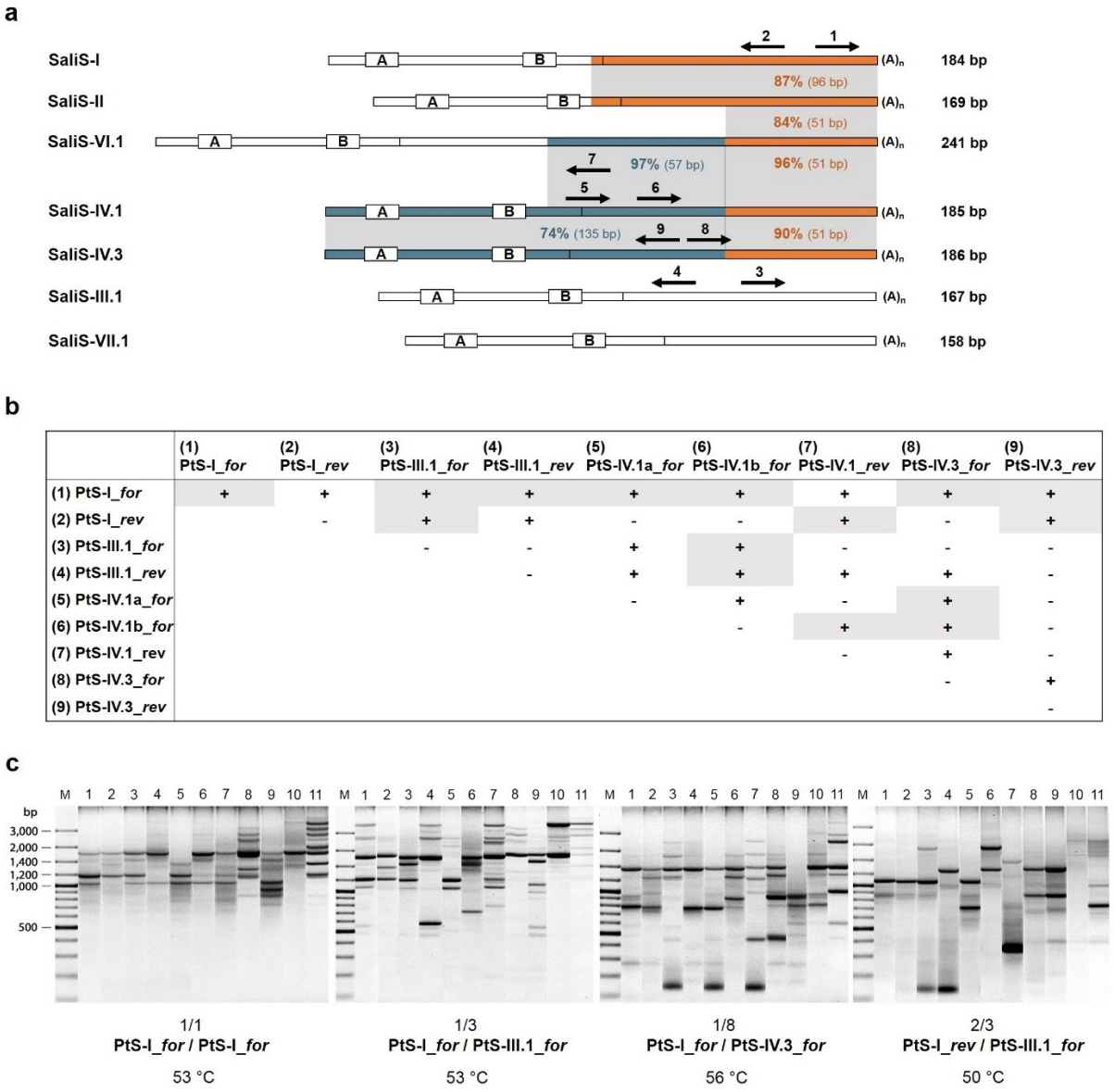


Figure 1. Development of ISAP markers for the differentiation of *P. tremula* genotypes. (a) The position and direction of *P. tremula* ISAP primers is indicated by black arrows on the respective SaliS families. A vertical line separates the SINE 5' region from the 3' region. The promotor motifs box A and box B are represented as boxes. Related SINE regions are shown by identical colors with the percentage similarities based on comparison of consensus sequences. (b) The 45 combinations of nine primers were tested with genomic DNA of *P. tremula* (accession 7590) as the PCR template. Primer combinations producing bands using the standard PCR (+) were tested under ISAP PCR conditions to identify primer pairs producing patterns containing more than four bands (grey shading). (c) Four informative primer pairs are shown for the following *P. tremula* genotypes: 1 - 10720-I, 2 - 10719-I, 3 - 10718-I, 4 - 10717-I, 5 - 10713-I, 6 - 10711-II, 7 - 10701-I, 8 - 10696-I, 9 - 10686-II, 10 - 7589, 11 - 7590. The respective annealing temperature is indicated below the gel images. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

The 45 primer combinations were tested with genomic DNA of *P. tremula* (accession 7590) in standard PCRs. Further optimization of the banding patterns was conducted with ISAP PCRs (Figure 1b). The primer PtS-I_*for* generates appropriate banding patterns with five of eight possible ISAP primers and additionally can be used as a single primer. The 13 primer combinations, producing more than four bands (Figure 1b), were subsequently applied to compare a collection of eleven *P. tremula* genotypes. The four primer pairs, which created most polymorphic bands are shown in Figure 1c.

Application of *P. tremula* ISAP primers in related Salicaceae species

Polymorphic ISAP markers, also enabling genotyping in related species, would substantially reduce time, costs and efforts for the selection of appropriate poplar clones for SRC cultivation. Since the SaliS families chosen for primer design are widely distributed in the genus *Populus* and also detectable in the willow species *Salix purpurea* (Chapter 2.3, Figure 1), the PtS primer set has the potential to be successfully applied for the discrimination of related poplars and hybrids thereof.

The ISAP profiles of *P. tremula* genotypes were registered using *GelComparII* to analyze the intraspecific genetic variability. Accessions of related species were added to the fingerprint database to enable interspecific genotype comparisons. A partial outcome including 30 poplar and ten willow accessions is shown in Figure 2.

The dendrogram represents the genetic similarity of the genotypes, also illustrated by grey scales (Figure 2a) and is based on the information of the ISAP banding patterns (Figure 2b). For each genotype, the bands resulting from five primer combinations were combined for the cluster analysis. The accessions of each species form separate branches in the dendrogram reflecting their closer genetic relationships (Figure 2a). The highest intraspecific diversity could be observed for *P. tremula*, while *P. trichocarpa*, and especially *P. nigra* genotypes, are presumably more similar to each other.

In *P. trichocarpa* and *P. nigra,* the *P. tremula*-derived ISAP primers produced fingerprints of sufficient quality for genotype comparisons (Figure 2b). However, not all primer combinations are suitable for this purpose, e.g. PtS-III.1_*rev* / PtS-IV.1b_*for* (4/6), producing only few bands in *P. nigra* and *P. trichocarpa* accessions (Figure 2b, lane 11 – 30).

The application of the PtS primers to the more distantly related genus *Salix* showed a strongly reduced number of bands, in particular for PtS-I_*for* / PtS-III.1_*rev* (1/4) and PtS-I_*for* / PtS-IV.3_*rev* (1/9) (Figure 2b, lane 31 – 40). Although *Salix alba* genotypes form a separate branch next to the *Salix fragilis* hybrids, the discrimination between *Salix* genotypes is presumably only possible to limited extent and requires *Salix*-specific ISAP primers.

The comparison of two genetically identical *P. tremula* accessions, 3110A and 3110B (Figure 2b, lane 3 – 4), showed identical ISAP profiles for four of five primer pairs. Using the primer pair PtS-I_*rev* / PtS-IV.3_*rev* (2/9), some bands are missing in 3110A compared to 3110B. Instead, a strong background smear was observed for 3110A, which is a typical result for either too high DNA concentrations or insufficient DNA purity. The DNA samples were extracted automatically (InnuPure C16 touch, Jena Analytik AG) without additional ethanol precipitations and optimization of the DNA quantity prior to the PCR (M. Brückner, personal communication).

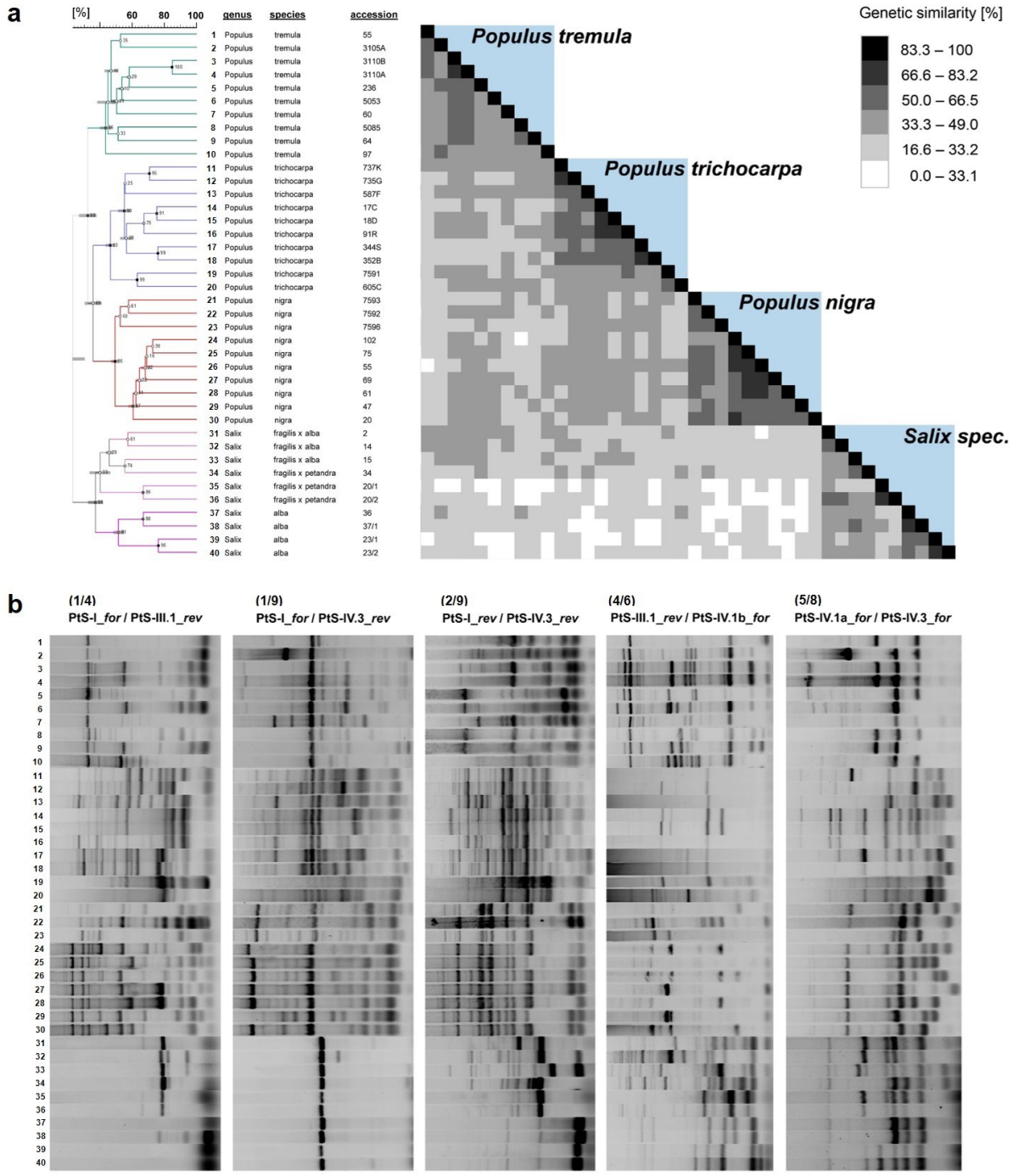


Figure 2. Application of *P. tremula* ISAP primers to genotypes of related species. (a) Genetic variability among ten genotypes each of the poplars *P. tremula* (green), *P. trichocarpa* (blue), and *P. nigra* (red) was compared with *S. alba* genotypes and *Salix* hybrids (pink) by UPGMA cluster analysis with the dice similarity coefficient for band matching. The percentage similarities are indicated as grey scales (see legend). The dendrogram branch quality was calculated with 1000 bootstrap simulations. (b) The individual fingerprints of all analyzed *Populus* (1-30) and *Salix* (31-40) genotypes based on five primer combinations. The data were compiled in cooperation with the Saxony State Forestry Service (Pirna, Germany).

Application of *P. tremula* ISAP primers for the differentiation of hybrid poplar clone accessions

A collection of 33 different hybrid poplar clones growing on a SRC testing area of the Saxony State Forestry Service was analyzed using a set of five *P. tremula* ISAP primer combinations. The collection mostly consists of *P. maximowiczii* × *P. trichocarpa*, but also of *P. maximowiczii* × *P. nigra* accessions (Figure 3). Each hybrid poplar clone is represented by five accessions each. Accordingly, the total number of 165 accessions was expected to form 33 cluster containing the five genetically identical individuals. However, the resulting UPGMA cluster analysis contained only 18 of 33 hybrid poplar clones showing the expected accession arrangement (not shown, M. Brückner, personal communication). A preference for correct accession arrangement related to the type of hybrid poplar clone, either *P. maximowiczii* × *P. trichocarpa* (mt) or *P. maximowiczii* × *P. nigra* (mn), was not observed. Two examples of the cluster analyses were provided by the Saxony State Forestry Service to illustrate correctly arranged accessions (Figure 3a) opposed to groups containing accessions of different hybrid poplar clones (Figure 3b).

The ISAP profiles of the 165 hybrid poplar accessions show 85 % - 98 % similarity (not shown, M. Brückner, personal communication). Although the number of detectable, polymorphic bands is relatively low for all primer pairs investigated, some of the ISAP primer combinations were more informative than others (Figure 3).

In an initial experiment, the primer pairs PtS-I_*for* /PtS-III.1_*rev* (1/4) and PtS-I_*for* / PtS-IV.3_*rev* (1/9) generated polymorphic ISAP profiles for the five clone pools (mt1 - mt4, mn1) investigated (Figure 3a). The primer combinations PtS-III.1_*rev* / PtS-IV.1b_*for* (4/6) and PtS-IV.1a_*for* / PtS-IV.3_*for* (5/8) show least bands and are presumably only suitable for the differentiation of *P. tremula* genotypes (Figure 2, lane 1 - 10; Figure 3). Using PtS-I_*rev* / PtS-IV.3_*rev* (2/9), informative banding patterns were generated, but less polymorphic: Especially mt2 - mt4 show highly similar ISAP profiles (Figure 3a). Nevertheless, the respective accessions of each hybrid poplar clone form a separate clade in the dendrogram.

However, the correct arragement of the clone pool accessions in the dendrogram depends on the total sample volume of the cluster analysis. This is demonstrated by the accessions of the poplar clone

‘mn1’, which are grouped together as expected in the initial experiment (Figure 3a, sample volume n = 25), but are placed separately in the complete cluster analysis (Figure 3b, sample volume n = 165). Also, for hybrid poplar clone pools showing highly similar ISAP profiles like mn1, mn2, and mn3, the tree topology does not correspond to the stated relation: The accessions of mn1 and mn2, respectively, are grouped to different clades, although arranged adjacently in the dendrogram (Figure 3b). In detail, the characteristic ISAP profile of the clone pool mn1 using PtS-I_*for* / PtS-IV.3_*rev* (1/9) shows that the two accessions ‘populus80’ and ‘populus115’ are part of a dendrogram clade of mixed accessions (Figure 3b), instead of being correctly arranged with the accessions ‘populus16’, ‘populus177’, and ‘populus33’.

Nevertheless, hybrid poplar clones with differing parent genomes were distinguished accordingly, indicated by separate main clades for the *P. maximowiczii* × *P. trichocarpa* and the *P. maximowiczii* × *P. nigra* clones, respectively (Figure 3).

The mn1 accession ‘populus115’ shows an inconsistent pattern using PtS-I_*for* / PtS-III.1_*rev* (1/4) and PtS-I_*rev* / PtS-IV.3_*rev* (2/9). Similar inconsistencies were observed for ‘populus112’ of mn2 using PtS-I_*for* / PtS-III.1_*rev* (1/4) and PtS-IV.1a_*for* / PtS-IV.3_*for* (5/8) (Figure 3b).

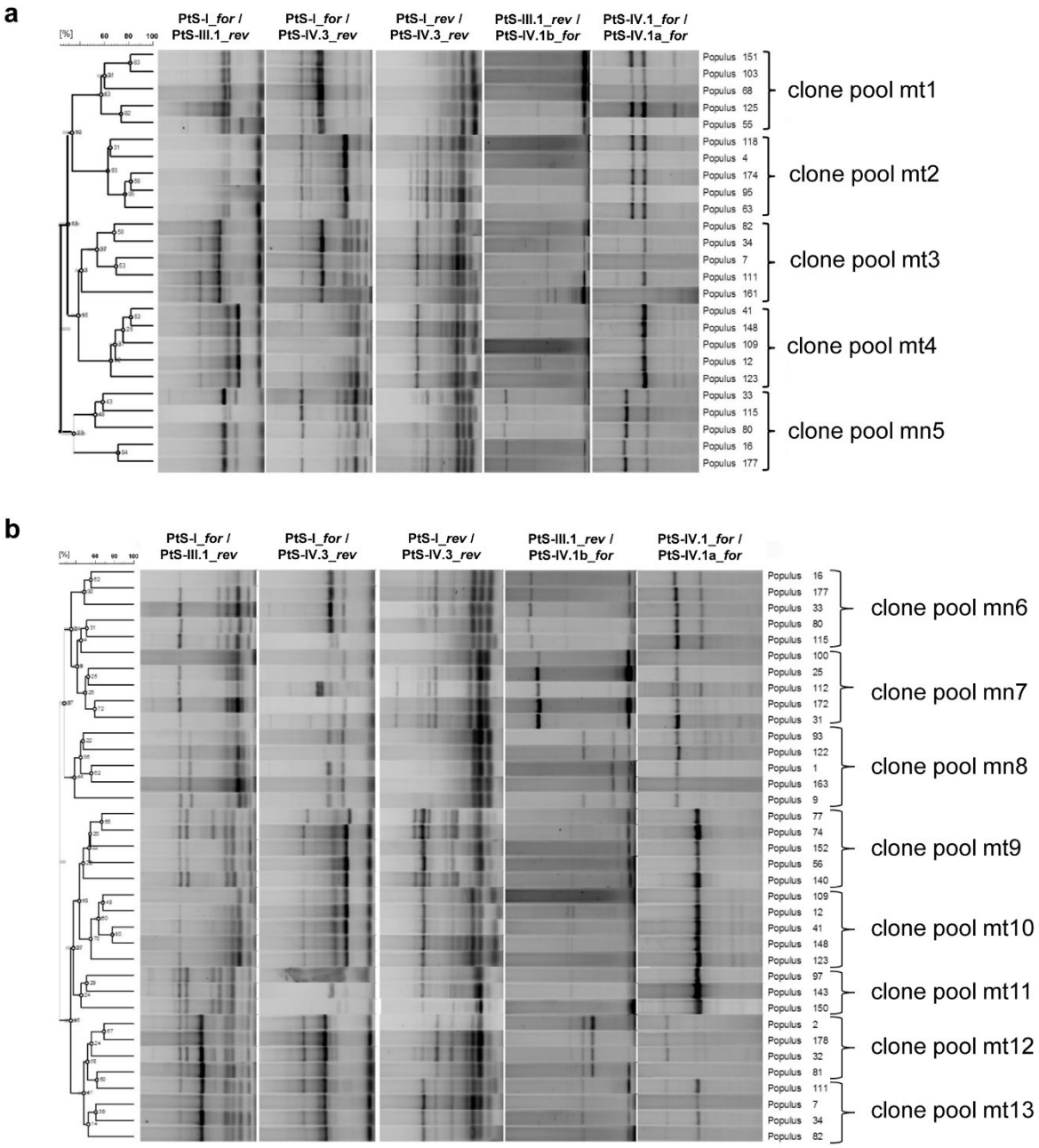


Figure 3. Genotyping of accessions from different hybrid poplar clones using *P. tremula* ISAP primers. Two examples of UPGMA cluster analyses using the dice similarity coefficient for band matching show *P. maximowiczii* × *P. trichocarpa* hybrids (mt) and *P. maximowiczii* × *P. nigra* hybrids (mn). The ISAP fingerprints are based on five *P. tremula* primer combinations. The dendrogram branch quality was calculated with 1000 bootstrap simulations. An intermediate result including 25 accessions shows examples of correct accession arrangement in (a), while an extract from the complete cluster analysis (33 hybrid poplar clones, 165 accessions) contains examples of incorrectly classified accessions (b). The data were provided by the Saxony State Forestry Service (Pirna, Germany).

Discussion

The genus *Populus* comprises 29 species, which are classified into six sections (Eckenwalder, 1996). The comparatively low species number is opposed to an extensive phenotypic variation within the *Populus* species. Frequently occurring natural interspecific hybridizations across the genus blur the species borders and often lead to taxonomic misclassification (Floate, 2004; Liesebach *et al.*, 2010; DiFazio *et al.*, 2011). Moreover, the intersectional cross-compatibility of *Populus* species offers considerable benefits for breeding, but constitutes a great challenge for the identification of involved species within hybrids (Tsarev *et al.*, 2016). Breeding institutions possess numerous promising hybrids whose original species contribution is not completely documented. However, for the registration of new poplar clones their genealogy has to be proven (Schroeder *et al.*, 2017).

Species-specific ISAP primers presumably only provide sufficient resolution for genotyping in closely related species

Based on the SINE families SaliS-I, SaliS-III, and SaliS-IV, nine ISAP primer were derived to establish the ISAP method for the European aspen (*Populus tremula*).

SaliS-I is highly abundant in *P. tremula* (Chapter 2.3, Figure 1; 1,174 full-length copies) and shares its 51 bp 3' end with other SaliS families (Figure 1a). Hence, the combination of both ISAP primers, PtS-I_*for* and PtS-I_*rev,* produced a smear on the agarose gel (not shown). However, applied as a single primer (PtS-I_*for*) and in combination with PtS-III.1_*for* and PtS-IV.3_*for*, the SaliS-I-derived primers contribute to the four most polymorphic banding patterns (Figure 1c). Thus, the special feature of the *P. tremula* ISAP primers is that they are able to bind to copies of several SINE families. Only SaliS-III-derived ISAP primers (PtS-III.1_*for* and PtS-III.1_*rev*) are family-specific (Figure 1a).

Although SINE families are usually scattered across plant families (Deragon and Zhang, 2006; Wenke *et al.*, 2011; Schwichtenberg *et al.*, 2016; Chapter 2.2, Figure 1; Chapter 2.3, Figure 1), ISAP primers only provide highest resolution exclusively for the species they were designed from.

The *P. tremula* ISAP primers were used to estimate the genetic variability of other *Populus* and *Salix* genotypes (Figure 2). ISAP patterns obtained for *P. trichocarpa* and *P. nigra* genotypes (Figure 2b)

contain numerous bands and hence, are considered as informative. However, the higher intraspecific genetic similarity of *P. trichocarpa*, and especially, of *P. nigra* genotypes (Figure 2a) might either reflect factual relationships or the insufficient resolution of the *P. tremula* ISAP markers in related *Populus* species. Presumably, the genetic differences between *P. trichocarpa* and *P. nigra* accessions, respectively, were not fully detected due to lower primer specificity, as the continuous evolutionary diversification and differentiation creates species-specific SINEs (Chapter 2.3, Figure 3). The application of *P. tremula* ISAP markers might still be reasonable in some close relatives, for example species of the same poplar section (Populus) like *Populus alba* or *Populus tremuloides.* This is supported by similar abundance, activity profiles and shared SINE subfamilies (SaliS-IV.3 and SaliS-VI.1) in *P. tremula* and *P. tremuloides* (Chapter 2.3, Figure 1, Figure 3).

The comparison of SINE copies in the monophyletic sister genera *Salix* and *Populus* (Wang *et al.*, 2014; Lauron-Moreau *et al.*, 2015), exemplified by *P. tremula* and *S. purpurea*, shows a species-specific differentiation of SaliS-IV.1 and SaliS-I (Chapter 2.3, Figure 3) and high sequence divergence like observed for SaliS-III.1 copies (Chapter 2.3, Figure 3). As extensive genome rearrangements accompanied the 'poplar-to-willow' process (Dickmann and Kuzovkina, 2008; Hou *et al.*, 2016), the evolutionary distance between *Salix* and *Populus* might impede a 'cross-genus' application of *P. tremula* ISAP primers while maintaining SINE-specific PCR products. The ISAP banding patterns of the more distantly related *Salix* species (Figure 2) might presumably contain a high proportion of amplicons resulting from random primer binding, which are less robust.

A collection of hybrid poplar clone accessions was comparatively analyzed using six SSR marker and five ISAP primer combinations. The 33 different *P. maximowiczii* hybrids of either *P. trichocarpa* or *P. nigra* could be identified with the established microsatellite markers WPMS09, WPMS12 (van der Schoot *et al.*, 2000) and WPMS18 (Smulders *et al.*, 2001) derived from *P. nigra* and PMGC456, PMGC2163, and PMGC2679 derived from *P. trichocarpa* (Poplar Molecular Genetics Cooperative, University of Washington, USA; http://www.ornl.gov/sci/ipgc/ssr_resource.html) (M. Brückner, personal communication).

The ISAP profiles within this experiment are generally less informative as observed in initial tests (Figure 2), probably due to the omitted purification of the DNA samples. The *P. tremula* ISAP primers

could not fully distinguish the total of 33 hybrid poplar clones (Figure 3). The genetic diversity of the clones, analyzed in a single UPGMA cluster analysis, determines the quantity of detectable clone pools:

The accessions propagated from a hybrid poplar clone can be identified by their arrangement in separate dendrogram clades (Figure 3a, e.g. clone pool mn1). However, the higher the genetic similarity among the clones, the higher the required number of polymorphic bands facilitating their differentiation / distinction of various clone pools (Figure 3a, mt2 – mt4). Hence, the 18 identified clone pools probably show a sufficient genetic diversity among each other to be distinguished by the species-unspecifically *P. tremula* ISAP primers.

A higher resolution might have been achieved using *P. maximowiczii* ISAP primers, as this poplar contributes to both poplar hybrids investigated. However, the SINE landscape of *P. maximowiczii* is not known and had to be analyzed. The *P. trichocarpa* SINEs also constitute suitable sources for an ISAP primer design. Especially the SINE family SaliS-I containing many evolutionarily young copies and the SaliS-IV.2 subfamily, which is probably still active (Chapter 2.3, Figure 3). Moreover, *P. trichocarpa* contributed to the intrasectional *P. maximowiczii* × *P. trichocarpa* hybrids and belongs to the same *Populus* section (Tacamahaca) like *P. maximowiczii* (DiFazio *et al.*, 2011) enabling also the differentiation of *P. maximowiczii* × *P. nigra* hybrids.

Genotyping based on the amplification of different microsatellite loci is also most effective, if species-specific primers are applied. However, they also produce reliable results for other poplar species (Rathmacher *et al.*, 2008; Bruegmann and Fladung, 2013). The cross-species transferability of *P. trichocarpa* SSR markers to *P. maximowiczii* genotypes was shown exemplified by the loci PMGC456 and PMGC2163 (Khasa *et al.*, 2005), while the applicability of *P. trichocarpa* microsatellite markers for the differentiation between aspens (*P. tremula*) and white poplars (*P. alba*) is only possible to limited extent (Yin *et al.*, 2009). AFLP analyses provided the same conclusion (Cervera *et al.*, 2005). This might be explained by the larger taxonomic distance, as *P. tremula* belongs to the *Populus* section Populus, while *P. trichocarpa* and *P. maximowiczii* are Tacamahaca species (Yin *et al.*, 2009; Liesebach *et al.*, 2010).

To determine the type of marker, ISAP or SSR, providing highest resolution for poplar cross-species applications, the genetic diversity of the investigated samples had to be characterized by highly informative sequencing-based markers like SNPs. Subsequently, the ISAP resolution might be evaluated and compared with the frequently used SSR markers according to standardized parameters (Nybom, 2004; Platten *et al.*, 2019).

The ISAP reveals indication for unstable poplar clone accessions

Vegetatively propagated poplar clone accessions originate from the same hybrid poplar specimen, forming a clone pool (Figure 3). Hence, they are expected to have the same genetic constitution.

However, inconsistent ISAP banding patterns of accessions derived from the same clone raise the question on the genetic stability of clone accessions. For example, the clone accession 'populus115' of the hybrid poplar specimen mn1 shows an altered ISAP profile for two of five primer pairs investigated compared to the four remaining clones (Figure 3b, PtS-I_*for* / PtS-III.1_*rev* and PtS-I_*rev* / PtS-IV.3_*rev*). Excluding sample contamination, a mutation of the specific ISAP primer binding site might have led to PCR products of altered length. In the context of SSR analyses, the 'loss' of one allele due to prevented primer binding is referred to as 'null allele' (Chapuis and Estoup, 2007). This points to the necessity of periodically inspections of the germplasm used for propagation to ensure the clone identity during long periods of clonal growth.

Other inconsistencies, associated with background smear, weak bands (Figure 3b, 'populus16' of mn1 analyzed with PtS-I_*for* / PtS-IV.3_*rev* (1/9)) or loss of bands (Figure 3b; 'populus97' of mt7 analyzed with PtS-I_*for* / PtS-III.1_*rev* (1/4)) result from insufficient DNA purity or inadequate DNA concentrations.

Microsatellite markers are still used for the differentiation between poplar clone collections (Ciftci *et al.*, 2017), although the tendency is towards complementary applications, like SSRs together with chloroplast SNPs (Schroeder *et al.*, 2017). However, even the combined application of AFLPs and SSRs, for example, could not guarantee the complete differentiation of the commercial poplar clones investigated, as twelve of 66 analyzed clones remained not distinguishable (Fossati *et al.*, 2005).

Furthermore, mitochondrial SNPs (Kersten *et al.*, 2015), SNPs from nuclear DNA (Mousavi *et al.*, 2016), and from RNAseq reads (Rogier *et al.*, 2018) as well as 5S rDNA-based marker (Alexandrov and Karlov, 2018) were applied and might be added to combined marker strategies.

ISAPs have the potential to contribute to sets of different molecular markers depending on the individual SINE properties in the genome of interest. Especially SINE families with evolutionarily young copies might contribute substantially for genotype differentiation, for example ISAP primer derived from SaliS-I and SaliS-IV.2 copies of *P. trichocarpa* (Chapter 2.3, Figure 3).

References

Alexandrov, S.O. and Karlov, I.G. (2018) Development of 5S rDNA-based molecular markers for the identification of *Populus deltoides* Bartr. ex Marshall, *Populus nigra* L., and their hybrids. *For.*, **9**, 604.

Baum, S., Weih, M., Busch, G., Kroiher, F. and Bolte, A. (2009) The impact of short rotation coppice plantations on phytodiversity. *Appl Agric For. Res*, **59**, 163–170.

Bruegmann, T. and Fladung, M. (2013) Potentials and limitations of the cross-species transfer of nuclear microsatellite marker in six species belonging to three sections of the genus *Populus* L. *Tree Genet. Genomes*, **9**, 1413–1421.

Cervera, M.T., Storme, V., Soto, A., Ivens, B., Montagu, M. Van, Rajora, O.P. and Boerjan, W. (2005) Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. *Theor. Appl. Genet.*, **111**, 1440–1456.

Chapuis, M.-P. and Estoup, A. (2007) Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.*, **24**, 621–631.

Ciftci, A., Karatay, H., Kücükosmanoğlu, F., Karahan, A. and Kaya, Z. (2017) Genetic differentiation between clone collections and natural populations of European black poplar (*Populus nigra* L.) in turkey. *Tree Genet. Genomes*, **13**, 69.

Dam, J. van, Faaij, A.P.C., Lewandowski, I. and Fischer, G. (2007) Biomass production potentials in Central and Eastern Europe under different scenarios. *Biomass and Bioenergy*, **31**, 345–366.

Dickmann, D.I. and Isebrands, J.G. (2001) Poplar clones: an introduction and caution. In Dickmann, D.I., Isebrands, J.G., Eckenwalder, J.E. and Richardson, J., eds. *Poplar culture in North America.* Part B, chapter 11. NRC Research Press, National Research Council of Canada, Ottawa, ON K1A0R6, Canada. pp 309-324.

Dickmann, D.I. and Kuzovkina, J. (2008) Poplars and willows of the world, with emphasis on silvicuturally important species. In *Poplar and willows in the world.* Rome, Italy: FAO.

DiFazio, S.P., Slavov, G.T. and Joshi, C.P. (2011) *Populus*: a premier pioneer system for plant genomics. In C. P. Joshi, S. P. DiFazio, and C. Kole, eds. *Genetics, genomics and breeding of poplar.* Enfield, NH: Science Publishers, pp. 1–28.

Dillen, S.Y., Djomo, S.N., Afas, N. Al, Vanbeveren, S. and Ceulemans, R. (2013) Biomass yield and

energy balance of a short-rotation poplar coppice with multiple clones on degraded land during 16 years. *Biomass and Bioenergy*, **56**, 157–165.

Eckenwalder, J.E. (1996) Systematics and evolution of *Populus*. In R. F. Stettler, H. D. Bradshaw Jr, P. E. Heilman, and T. M. Hinckley, eds. *Biology of Populus and its implications for management and conservation.* Ottawa, ON, Canada: NRC Research Press, pp. 7–32.

Eppler, U., Petersen, J.-E. and Couturier, C. (2007) *Short rotation forestry, short rotation coppice and perennial grasses in the European Union: agro-environmental aspects, present use and perspectives* J. F. Dallemand, J. E. Petersen, and A. Karp, eds., Harpenden, United Kingdom: European Commission, Joint Research Centre, Institute for Energy, Renewable Energy Unit.

Floate, K.D. (2004) Extent and patterns of hybridization among the three species of *Populus* that constitute the riparian forest of southern Alberta, Canada. *Can. J. Bot.*, **82**, 253–264.

Fossati, T., Zapelli, I., Bisoffi, S., Micheletti, A., Vietto, L., Sala, F. and Castiglione, S. (2005) Genetic relationships and clonal identity in a collection of commercially relevant poplar cultivars assessed by AFLP and SSR. *Tree Genet. Genomes*, **1**, 11–20.

Heräjärvi, H. and Junkkonen, R. (2006) Wood density and growth rate of European and hybrid aspen in Southern Finland. *Balt. For.*, **12**, 2–8.

Hou, J., Ye, N., Dong, Z., Lu, M., Li, L. and Yin, T. (2016) Major chromosomal rearrangements distinguish willow and poplar after the ancestral "salicoid" genome duplication. *Genome Biol. Evol.*, **8**, 1868–1875.

Kersten, B., Voss, M.M. and Fladung, M. (2015) Development of mitochondrial SNP markers in different *Populus* species. *Trees*, **29**, 575–582.

Khasa, D., Pollefeys, P., Navarro-Quezada, A., Perinet, P. and Bousquet, J. (2005) Species-specific microsatellite markers to monitor gene flow between exotic poplars and their natural relatives in eastern North America. *Mol. Ecol. Notes*, **5**, 920–923.

Lasch, P., Kollas, C., Rock, J. and Suckow, F. (2010) Potentials and impacts of short-rotation coppice plantation with aspen in Eastern Germany under conditions of climate change. *Reg. Environ. Chang.*, **10**, 83–94.

Lauron-Moreau, A., Pitre, F.E., Argus, G.W., Labrecque, M. and Brouillet, L. (2015) Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS One*, **10**, e0138963.

Leibundgut, H. (1967) Pappeln als Baumarten des Vorwaldes. *Mitteilungen der Schweizerische Pappel-Arbeitsgemeinschaft*, **12**, 1–7.

Liesebach, H., Schneck, V. and Ewald, E. (2010) Clonal fingerprinting in the genus *Populus* L. by nuclear microsatellite loci regarding differences between sections, species and hybrids. *Tree Genet. Genomes*, **6**, 259–269.

Liesebach, M. ed. (2015) *FastWOOD II: Züchtung schnell- wachsender Baumarten für die Produktion nachwachsender Rohstoffe im Kurzumtrieb – Erkenntnisse aus 6 Jahren FastWOOD* [In German], Braunschweig: Johann Heinrich von Thünen-Institut.

Liesebach, M., Wuehlisch, G. von and Muhs, H.-J. (1999) Aspen for short-rotation coppice plantations on agricultural sites in Germany: effects of spacing and rotation time on growth and biomass production of aspen progenies. *For. Ecol. Manage.*, **121**, 25–39.

Lin, Y.-C., Wang, J., Delhomme, N., et al. (2018) Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proc. Natl. Acad. Sci.*, **115**, E10970–E10978.

Mohrdiek, O. (1977) Hybridaspen für forstliche Grenzertragsböden. *Forstarchiv*, **48**, 158–163.

Mousavi, M., Tong, C., Liu, F., Tao, S., Wu, J., Li, H. and Shi, J. (2016) *De novo* SNP discovery and genetic linkage mapping in poplar using restriction site associated DNA and whole-genome sequencing technologies. *BMC Genomics*, **17**, 656.

Niemczyk, M., Wojda, T. and Kaliszewski, A. (2016) Biomass productivity of selected poplar (*Populus* spp.) cultivars in short rotations in northern Poland. *New Zeal. J. For. Sci.*, **46**, 22.

Nybom, H. (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.*, **13**, 1143–1155.

Pearce, D.W., Zanewich, K.P. and Rood, S.B. (2018) Heterosis in poplar involves phenotypic stability: cottonwood hybrids outperform their parental species at suboptimal temperatures. *Tree Physiol.*, **38**, 789–800.

Platten, J.D., Cobb, J.N. and Zantua, R.E. (2019) Criteria for evaluating molecular markers: comprehensive quality metrics to improve marker-assisted selection. *PLoS One*, **14**, e0210529.

Pontailler, J., Ceulemans, R. and Guittet, J. (1999) Biomass yield of poplar after five 2-year coppice rotations. *For.*, **72**, 157–163.

Rathmacher, G., Niggemann, M., Wypukol, H., Gebhardt, K., Ziegenhagen, B. and Bialozyt, R. (2008) Allelic ladders and reference genotypes for a rigorous standardization of poplar microsatellite data. *Trees*, **23**, 573.

Rogier, O., Chateigner, A., Amanzougarene, S., et al. (2018) Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*. *BMC Genomics*, **19**, 909.

Schoot, J. van der, Pospíšková, M., Vosman, B. and Smulders, M.J.M. (2000) Development and characterization of microsatellite markers in black poplar (*Populus nigra* L.). *Theor. Appl. Genet.*, **101**, 317–322.

Schroeder, H., Kersten, B. and Fladung, M. (2017) Development of multiplexed marker sets to identify the most relevant poplar species for breeding. *For.*, **8**, 492.

Sixto, H., Gil, P., Ciria, P., Camps, F., Sánchez, M., Cañellas, I. and Voltas, J. (2014) Performance of hybrid poplar clones in short rotation coppice in Mediterranean environments: analysis of genotypic stability. *GCB Bioenergy*, **6**, 661–671.

Smulders, M.J.M., Schoot, J. Van Der, Arens, P. and Vosman, B. (2001) Trinucleotide repeat microsatellite markers for black poplar (*Populus nigra* L.). *Mol. Ecol. Notes*, **1**, 188–190.

Tsarev, A., Wühlisch, G. von and Tsareva, R. (2016) Hybridization of poplars in the central Chernozem region of Russia. *Silvae Genet.*, **65**, 1–10.

Vanbeveren, S.P.P., Spinelli, R., Eisenbies, M., Schweier, J., Mola-Yudego, B., Magagnotti, N., Acuna, M., Dimitriou, I. and Ceulemans, R. (2017) Mechanised harvesting of short-rotation coppices. *Renew. Sustain. Energy Rev.*, **76**, 90–104.

Wang, Z., Du, S., Dayanandan, S., Wang, D., Zeng, Y. and Zhang, J. (2014) Phylogeny reconstruction and hybrid analysis of populus (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. *PLoS One*, **9**, e103645.

Yemshanov, D. and McKenney, D. (2008) Fast-growing poplar plantations as a bioenergy supply source for Canada. *Biomass and Bioenergy*, **32**, 185–197.

Yin, T.M., Zhang, X.Y., Gunter, L.E., Li, S.X., Wullschleger, S.D., Huang, M.R. and Tuskan, G.A. (2009) Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. *New Phytol.*, **181**, 498–503.

3.3 Evaluation of the genetic composition of *Larix* hybrids (*Larix* × *eurolepis*) for the targeted identification of economically valuable phenotypes

Introduction

Fast-growing tree species are required in forestry. As an alternative to the commonly planted coniferous species in Europe (e.g. Norway spruce, Douglas fir), the hybrid larch (*Larix* × *eurolepis* Henry) gains importance (Pâques *et al.*, 2013). Interspecific hybrids of European (*Larix decidua* Mill.) and Japanese larch (*Larix kaempferi* (Lamb.) Carr.) emerged at the beginning of the 20th century in Scotland (Henry and Flood, 1919). Compared to the parent species, these hybrid conifers exhibit heterosis in growth performance and stem form (Matyssek and Schulze, 1987; Eko *et al.*, 2004; Pâques, 2009; Marchal *et al.*, 2017). The high durability of larch wood is suitable for outdoor uses (e.g. boat building, fence posts, garden furniture) and represents a green alternative to impregnated wood (Larsson-Stern, 2003; Pâques *et al.*, 2013).

However, despite the superior economic potential, cultivated areas remained small (Perks *et al.*, 2006), primarily due to the challenging vegetative and generative reproduction: poor seed production and low germination capacity (Lelu *et al.*, 1994) decelerates progress in breeding, and the poor efficiency of conventional vegetative propagation (e.g. cuttings) impedes the mass production of proven varieties (Harrison, 2002; Perks *et al.*, 2006). The most effective technique for clonal propagation of hybrid larches is somatic embryogenesis (Klimaszewska, 1989; Lelu-Walter and Pâques, 2009; Kraft and Kadolsky, 2018). Another critical aspect of hybrid larch breeding is the highly variable genetic constitution of the progeny: open-pollinated seed orchards from European and Japanese larch hybridizations contain high proportions of parent genotypes, but only little hybrid character (Lee, 2003).

Within the TreeSINE consortium the applicability of the ISAP marker system for the characterization of genome components in hybrid larch progeny was intended to be examined. Parent species and F1 offspring genotypes have to be comparatively analyzed with respect to the ratio of *L. decidua*- and *L. kaempferi*-specific bands.

Experimental procedures

Plant material and DNA preparation

Genomic DNA of the European larch (*Larix decidua*) was extracted from lyophilized needles of a reference specimen growing at the Forest Botanical Garden of Tharandt (genotype 'Tharandt'; GPS coordinates Lat 50.98279 and Lon 13.57901). The genomic DNA of the 'Tharandt' genotype was extracted using the 'DNeasy Plant Maxi Kit' (Qiagen, Valencia, USA) and was used for Illumina sequencing and initial ISAP / ISRAP analyses. Genomic DNA of six *Larix decidua* accessions with the breeding numbers '91' and '45' and the seed numbers '43 (36)', '10 (10)', '48 (366)', '6 (6)' were obtained by the Saxony State Forestry Service (Pirna, Germany). The DNA quality control was conducted as described in Chapter 3.1.

DNA sequencing

Next Generation Sequencing (NGS) data of the *L. decidua* genotype 'Tharandt' were generated by Macrogen Inc. (Seoul, South Korea) using Illumina technology. Two sequence libraries with different insert sizes were sequenced in paired-end mode (Table 1).

Table 1. Characteristics of the *Larix decidua* (genotype 'Tharandt') sequence libraries.

Sequence library	**[1]**	**[2]**
Illumina sequencing system	HiSeq 4000	HiSeq 2000
Insert size ordered [bp]	180	350
Insert size received [bp]	300	470
Read length [bp]	101	101
Read count	2 x 418,320,282	2 x 217,781,121
Size (Gb)	~ 84.5	~ 44.0

SINE identification

Basically, the identification of *L. decidua* SINEs and the family classification was carried out as described for *C. japonica* SINEs in Chapter 2.1. However, comprehensive genomic sequence data were not available in public sequence databases (year 2016) and an assembly of satisfactory quality could not be achieved due to the absence of long sequencing reads and low computational capacity regarding the genome size of 13 Gb (Zonneveld, 2012). Therefore, an adapted approach was developed: the SINE identification based on 101 bp paired-end Illumina reads (Table 1) required a preprocessing of the sequencing data to enable the application of the *SINE-Finder* and a modified procedure to detect the SINE family consensus sequences (Figure 1).

The 101 bp forward and reverse reads were concatenated and analyzed with the *SINE-Finder* (Figure 1a).

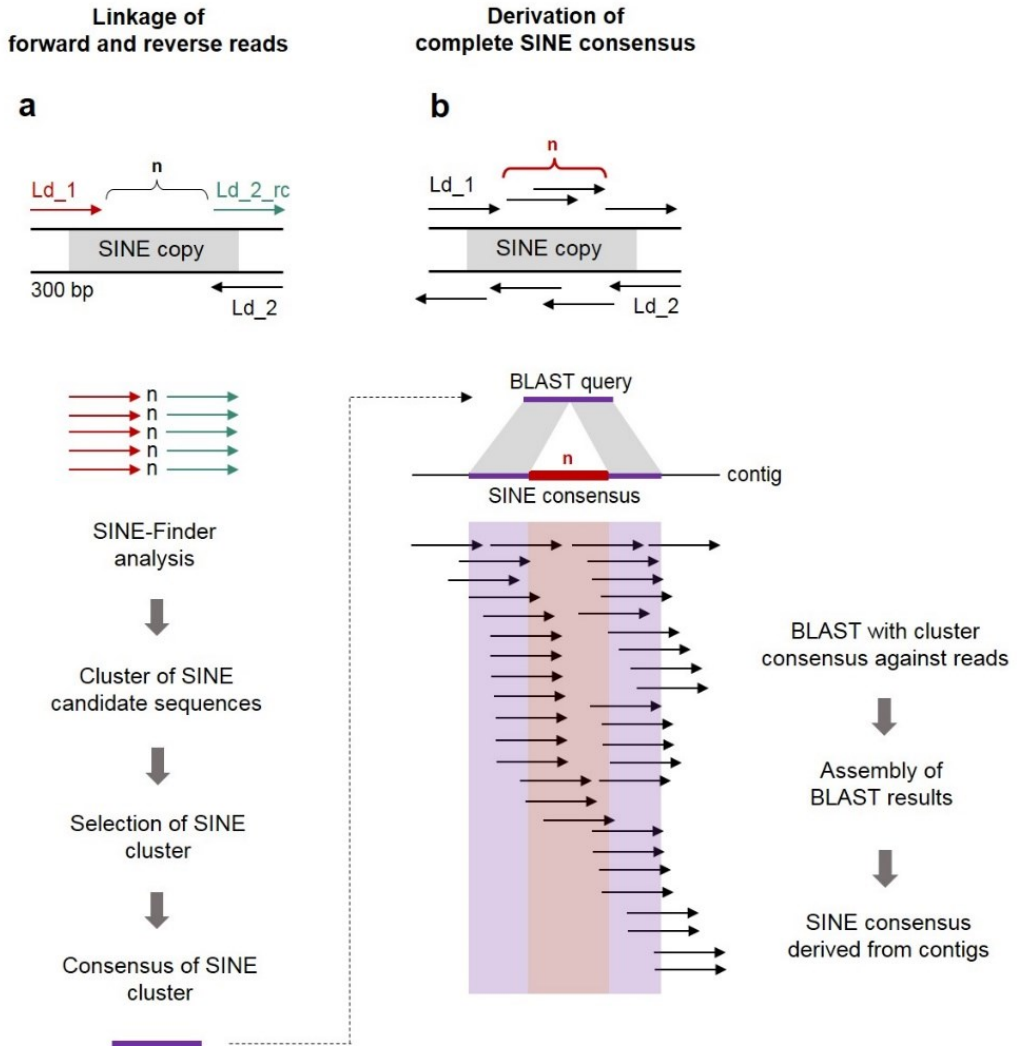


Figure 1. Principle of the SINE identification based on Illumina raw reads. (a) Exemplary, a 300 bp insert of a sequencing library, containing a SINE, is represented with the resulting sequencing reads (arrows in a and b). The reverse reads (Ld_2, black arrows) were translated to reverse complementary orientation (Ld_2_rc, green arrows) and concatenated with the forward reads (Ld_1, red arrows). These read constructs statistically contain all structural features of a SINE copy necessary for the detection by the SINE finder. (b) Consensus sequences derived from each SINE cluster (purple bar, a and b) were used as queries for *BLAST* searches in a database composed of the sequencing raw reads (Ld_1 and Ld_2, black arrows in b). The *BLAST* output reads were *de novo* assembled and the resulting contigs were analyzed to complete the SINE consensus sequence by the mid region 'n' (red).

In case of a central SINE position on the respective insert (Figure 1a), these sequence constructs contain all necessary SINE features for a detection by the *SINE-Finder*. The yet unknown sequence of the central SINE region is represented by the letter 'n', connecting the paired-end reads of an insert. To determine the missing central SINE region 'n', the consensus sequence of each SINE cluster (Figure 1a, purple) was used as query for *BLAST* (Altschul *et al.*, 1990) searches in the sequencing reads (Figure 1b). The *de novo* assembly of the resulting *BLAST* hits provides contigs, which are artificial sequences originating from read overlaps (Figure 1b). These contigs contain the complete SINE consensus sequence, which can be recognized by a region of sequence conservation within the assembled reads (Figure 1b, purple and red shaded regions). This region is terminated by poly (A) stretches of different length and flanked by variable regions, corresponding to the different genomic environment of each SINE copy. To obtain more representative SINE family consensus sequences, the initial complete SINE consensus was used for a *BLAST* in the read database and the resulting hits were mapped to the search query. The refined complete SINE family consensus sequences were used for the ISAP primer design.

ISAP PCR and agarose gel electrophoresis

The ISAP experiments were carried out as described in Chapter 3.1. The ISAP primers derived from the PinS families of *L. decidua* are listed in Table 2.

Table 2. ISAP primer. For standard PCR the 20mer primers were used as listed. For the ISAP PCR the SINE-derived primers were extended by a 5' GC-rich extension (5' - CTGACGGGCCTAACGGAGCG - 3') resulting in 40mer primers.

SINE family	***forward* Primer**		***reverse* Primer**	
	name	sequence (5' - 3' orientation)	name	sequence (reverse complement)
PinS-II	LdS-II_*for*	CTTGGGAGGTTGTTGTTCCC	LdS-II_*rev*	ACTTGTGACTCAGCAGGGGC
PinS-III	LdS-III_*for*	TTCGGAATAGCAGGAAGGTG	LdS-III_*rev*	TCGAGCAAACCGTCAGCCGG
PinS-VI	LdS-VI_*for*	CCATTGAGCGCCGGTTWCAC	LdS-VI_*rev*	AATCGGACGGGGTCTCGGGG

Preparation of a DNA restriction fragment library for ISRAP assays

As a basis for an inter-SINE-restriction site amplified polymorphism (ISRAP) assay, a DNA restriction fragment library was created for each *L. decidua* genotype. Genomic DNA was digested with the restriction endonuclease *Eco*RI (Thermo Scientific, Waltham, USA) and the cleavage sites were capped with respective *Eco*RI adapters (Table 3).

Table 3. *Eco*RI adapter and *Eco*RI adapter primer. Sequences for adapter hybridization and *Eco*RI adapter primer with three (GAC) and two (GA) selective nucleotides are listed. *Eco*RI adapter primers were extended by a 5' GC-rich extension (5' – CTGACGGGCCTAACGGAGCG – 3') for ISAP and ISRAP applications. For the fragment length analysis, *Eco*RI adapter primers were labeled with ATTO550 according to the Eurofins Dye Set EF-01 (Eurofins Genomics, Ebersberg, Germany).

name	**sequence (5' - 3')**
*Eco*RI-adapter1	CTCGTAGACTGCGTACC
*Eco*RI-adapter2	TTAACCATGCGTCAGATG
*Eco*RI-F-GAC_ext	GACTGCGTACCAATTCGAC
*Eco*RI-F-GA_ext	GACTGCGTACCAATTCGA
*Eco*RI-F-GAC_A550	GACTGCGTACCAATTCGAC

The *Eco*RI adapters were hybridized from the corresponding oligonucleotides (Table 3) by incubation at 94 °C for 3 minutes and cooling down to room temperature:

Reaction mix:

Distilled water	180.0 µl
*Eco*RI-adapter1 (100 mM)	10.0 µl
*Eco*RI-adapter2 (100 mM)	10.0 µl
Total volume	200.0 µl

Genomic DNA was digested with the restriction endonuclease *Eco*RI according to the following scheme:

Reaction mix:

Genomic DNA (1 µg)	x µl
Distilled water	y µl
10× *Eco*RI buffer	8.0 µl
Restriction endonuclease *Eco*RI (10 U/µl)	1.0 µl
Total volume	80.0 µl

Distilled water (volume 'y') was added to the mixture of DNA (volume 'x'), buffer and *Eco*RI to reach the total reaction volume of 80 µl. The reaction mix was incubated for one hour at 37 °C and purified with the 'GeneJet Gel Extraction and DNA Cleanup Micro Kit' (Protocol A - General DNA Cleanup from enzymatic reactions, Thermo Scientific, Waltham, USA). Deviating from the standard procedure, the DNA fragments were eluted twice with 20 µl of distilled water each.

The ligation of the *Eco*RI adapters to the *Eco*RI-digested DNA was performed at 10 °C overnight:

Reaction mix:

*Eco*RI-digested DNA	x µl
Distilled water	y µl
10× T4 DNA ligase buffer	5.0 µl
*Eco*RI adapter (5 mM)	2.0 µl
T4 DNA ligase (3 U/µl)	1.0 µl
Total volume	50.0 µl

The enzymatic reaction of the DNA T4 ligase (Promega, Madison, USA) was stopped by incubation at 65 °C for 10 minutes in the ThermoMixer (Eppendorf, Hamburg, Germany).

For verification of ligation, a PCR assay (standard PCR program, 50 °C annealing temperature) was conducted using the *Eco*RI adapter primers (Table 3) and the ligation reaction mix as DNA template (after enzyme reaction). Based on the assumption that the *L. decidua* genome does not contain the *Eco*RI adapter sequence, the respective 1,2 % agarose gel was expected to show a smear due to the amplification of the *Eco*RI-digested DNA. Genomic *L. decidua* DNA was tested with the *Eco*RI adapter primers as negative control. The amplification reaction was carried out in a final volume of 20 µl as follows:

PCR ingredients:

Distilled water	10.7 µl
10× DreamTaq™ Green Buffer	2.0 µl
dNTPs (2 mM)	2.0 µl
BSA (bovine serum albumin) (2 mg/ml)	2.0 µl
*Eco*RI adapter primer (10 µM)	2.0 µl
Ligation reaction mix	1.0 µl
DreamTaq™ DNA polymerase (5 U/µl)	0.3 µl
Total volume	20.0 µl

Standard PCR:

94 °C	5 min		initial denaturation
94 °C	20 s	30 ×	denaturation
50 °C	30 s		annealing
72 °C	2 min		extension
72 °C	5 min		final extension
4 °C	∞		storage

Subsequently, the *Eco*RI-digested and *Eco*RI adapter-capped DNA fragments were digested with the restriction endonuclease *Mse*I (Thermo Scientific, Waltham, USA):

Reaction mix:

*Eco*RI-digested DNA	x µl
Distilled water	y µl
10× buffer R	8.0 µl
Restriction endonuclease *Mse*I (10 U/µl)	1.0 µl
Total volume	80.0 µl

The reaction mix was incubated for one hour at 65 °C. The DNA fragments were purified with the 'GeneJet Gel Extraction and DNA Cleanup Micro Kit' (Thermo Scientific, Waltham, USA) and eluated twice with 20 µl of distilled water.

Genomic DNA treated according to this procedure is designated 'DNA restriction fragment library' and can be used as DNA template for ISRAP PCR assays.

ISRAP PCR and capillary gel electrophoresis

In the ISRAP PCRs up to three different SINE-derived primers (ISAP primers) were combined with the *Eco*RI adapter primer. The 'DNA restriction fragment library' of the genotype of interest was added as DNA template. The reaction mix was prepared as follows:

PCR ingredients:

DNA restriction fragment library (~ 20 ng/µl)	1.0 µl
Distilled water	8.7 - 10.7 µl
10× DreamTaq™ Green Buffer	2.0 µl
dNTPs (2 mM)	2.0 µl
BSA (bovine serum albumin) (2 mg/ml)	2.0 µl
*Eco*RI adapter primer (10 µM)	1.0 µl
1 – 3 SINE-derived primer (10 µM)	1.0 – 3.0 µl
DreamTaq™ DNA polymerase (5 U/µl)	0.3 µl
Total volume	20.0 µl

ISRAP PCR program:

93 °C	5 min		initial denaturation
93 °C	20 s	3 ×	denaturation
50 °C	30 s		annealing
72 °C	2 min		extension
93 °C	20 s	27 ×	denaturation
72 °C	140 s		annealing/extension
72 °C	5 min		final extension
4 °C	∞		storage

The gel images of the *L. decidua* ISRAP (and ISAP) profiles were captured with the VWR® Imager (VWR International GmbH, Radnor, USA).

For a separation of ISRAP PCR products by capillary gel electrophoresis, *Eco*RI adapter primer labeled with the fluorescent dye ATTO550 (Eurofins Genomics, Ebersberg, Germany) were used (Table 3). One half of the total reaction volume was separated by conventional agarose gel electrophoresis and the remaining 10 µl were sent to Eurofins Genomics (Ebersberg, Germany) for a fragment length analysis (FLA) using the ABI 3130 XL sequencing machine.

Results

The establishment of the ISAP marker system for *L. decidua* genotype comparisons reveals an insufficient polymorphism count

The SINE identification based on Illumina sequencing reads provides consensus sequences, each representing a SINE family (Figure 1). The European larch (*Larix decidua* Mill.) contains the six Pinaceae SINE (PinS) families PinS-I to PinS-VI (Figure 2a).

Previously, several PinS-I copies were detected in loblolly pine (*Pinus taeda*), white spruce (*Picea glauca*) and Sitka spruce (*Picea sitchensis*) (Wenke *et al.*, 2011). They differ from the PinS-I copies identified from the *L. decidua* short sequencing reads (Table 1), in particular in their 3' regions. However, sequence comparisons revealed 86 % similarity over 107 bp beginning at the 5' end (Figure 2a) and an overall similarity of 66 %, indicating a subfamily structure. Hence, the previously reported PinS-I SINEs (Wenke *et al.*, 2011) were designated PinS-I.1, while the PinS-I copies identified in *L. decidua* were designated PinS-I.2. Subsequently, PinS-I.1 could also be detected in the *L. decidua* sequencing reads by *BLAST* searches using the published consensus sequence (Wenke *et al.*, 2011).

The PinS-I SINE families could not be characterized regarding abundance and similarity, as the *L. decidua* sequencing reads are too short to contain full-length SINEs. Thus, ISAP primers, designated LdS (*Larix decidua* SINE), were derived from the PinS families with the highest read count of the *SINE-Finder* output: PinS-II, PinS-III and PinS-VI (Figure 2a, Table 3). PinS-VI constitutes a candidate SINE family to derive informative ISAP primers, since it contains a high number of *SINE-Finder* output sequences. However, it is the only SINE family of short length (126 bp), while the remaining PinS families are over 200 bp in length. The small 3' region of PinS-VI does not offer the possibility to derive forward and reverse ISAP primer of differing sequence. Thus, the primer LdS-VI_*rev* contains large parts of the SINE 5' region (Figure 2a), however, without including the highly conserved nucleotides of the 11 bp box B motif (GGTTCGAnnCC; Galli *et al.*, 1981).

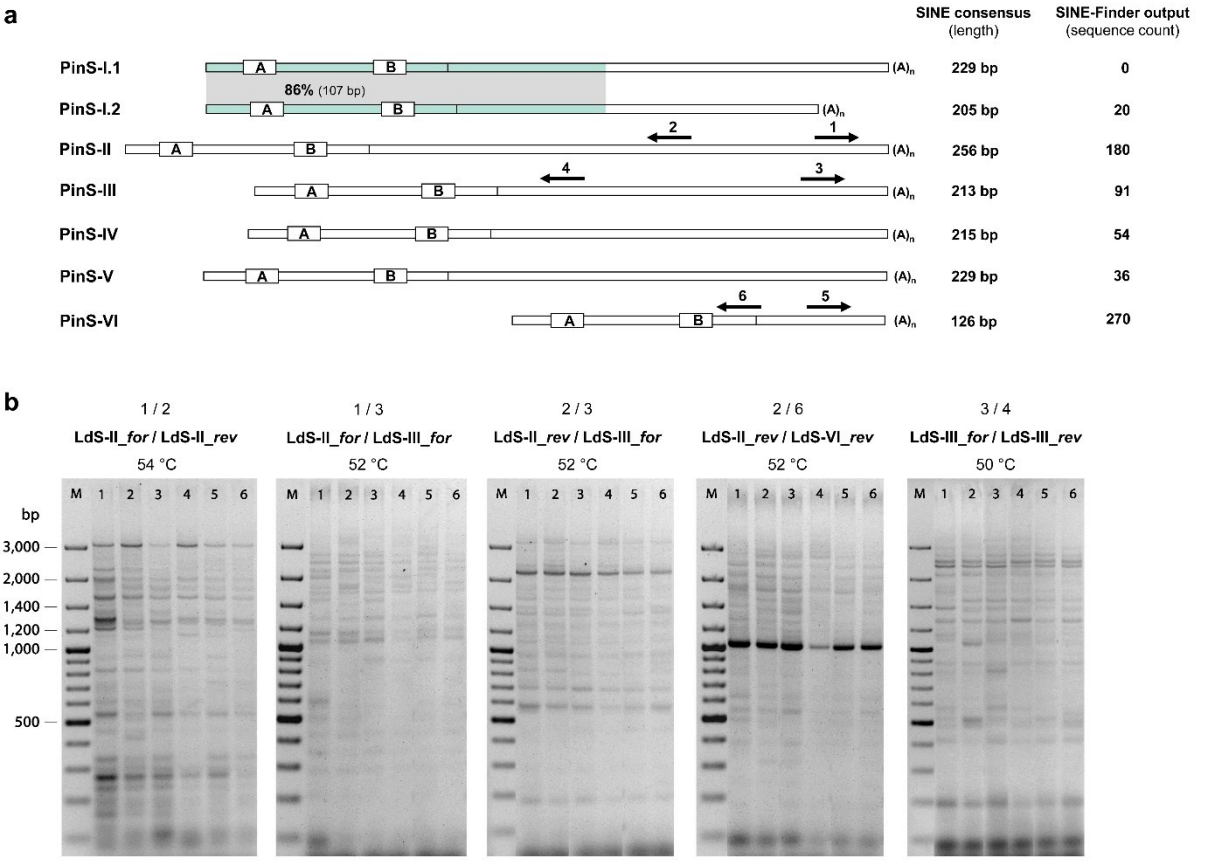


Figure 2. Analysis of *L. decidua* ISAP primer combinations reveals low information content. (a) The *L. decidua* ISAP primers are represented by black arrows showing position and direction on the PinS families (consensus sequences): 1 – LdS-II_*for*, 2 – LdS-II_*rev*, 3 – LdS-III_*for*, 4 – LdS-III_*rev*, 5 – LdS-VI_*for*, 6 – LdS-VI_*rev*. The SINE 5' region is separated from the 3' region by a vertical line. The promotor motifs box A and box B are represented as boxes. The related region of the PinS-I subfamilies is shown by identical color with the respective percentage similarity. (b) The LdS primer combinations generating numerous bands were applied to *L. decidua* genotypes obtained by the Saxony State Forestry Service: 1 - breeding number 91, 2 - breeding number 45, 3 - seed number 43 (36), 4 - seed number 10 (10), 5 - seed number 48 (366), 6 - seed number 6 (6). The ISAP primer combinations, including their respective annealing temperatures, are indicated above the gel images. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

The ISAP profiles of the *L. decidua* genotypes show a high density of bands (Figure 2b). However, the band intensity is too low for an automated evaluation of gel images using *BioNumerics* and, moreover, polymorphisms are rare.Regarding the highly similar patterns in different genotypes (Figure 2b), the LdS ISAP primers are not feasible for the discrimination of highly related *L. decidua* genotypes. Four F1 offspring genotypes (Figure 2b, 3-6) were compared with the respective crossing parents (Figure 2b, 1-2). Only three of five primer combinations investigated show slight differences between the banding patterns of the crossing parents: LdS-II_*for* / LdS-II_*rev*, LdS-II_*for* / LdS-III_*for*, and LdS-III_*for* / LdS-III_*rev.* Thus, alternative polymorphic genome loci like unequally distributed restriction sites have to be included resulting in a combined ISAP and AFLP technique.

The inter-SINE-restriction site amplified polymorphism (ISRAP) - Development of a marker system combining ISAP and AFLP technique

The inter-SINE-restriction site amplified polymorphism (ISRAP) method combines SINE-derived primers (ISAP primers) with primers specifically binding to *Eco*RI adapter sequences (Figure 3; Table 3, *Eco*RI adapter primers).

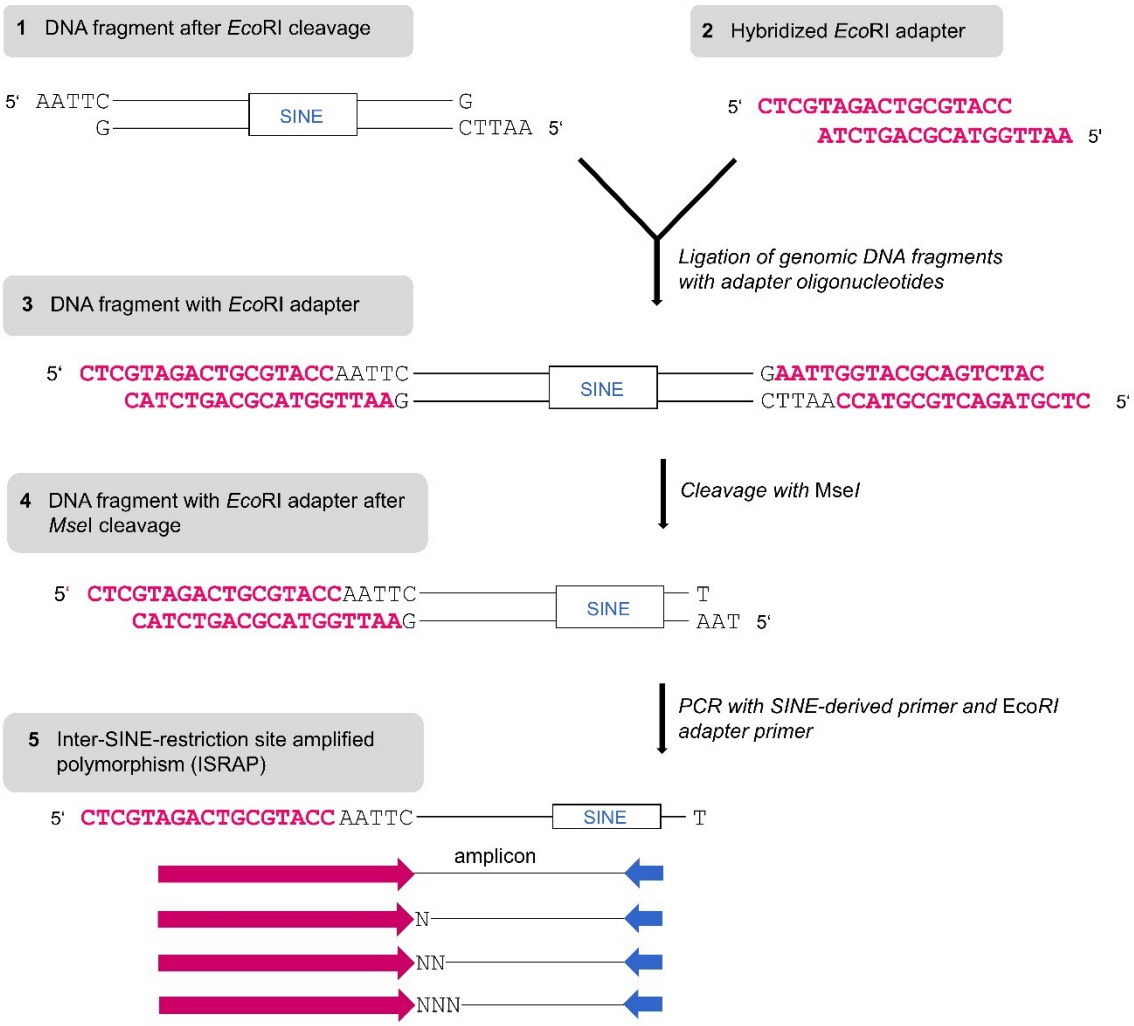


Figure 3. Principle of the ISRAP marker system. *Eco*RI-digested genomic DNA of the European larch is ligated with respective adapter molecules (1-3). Subsequent *Mse*I-cleavage (4) increases the fraction of amplicons based on SINEs and *Eco*RI cleavage sites (5). Up to three selective nucleotides (N) at the 3' end of the *Eco*RI adapter primer regulate the number of PCR products.

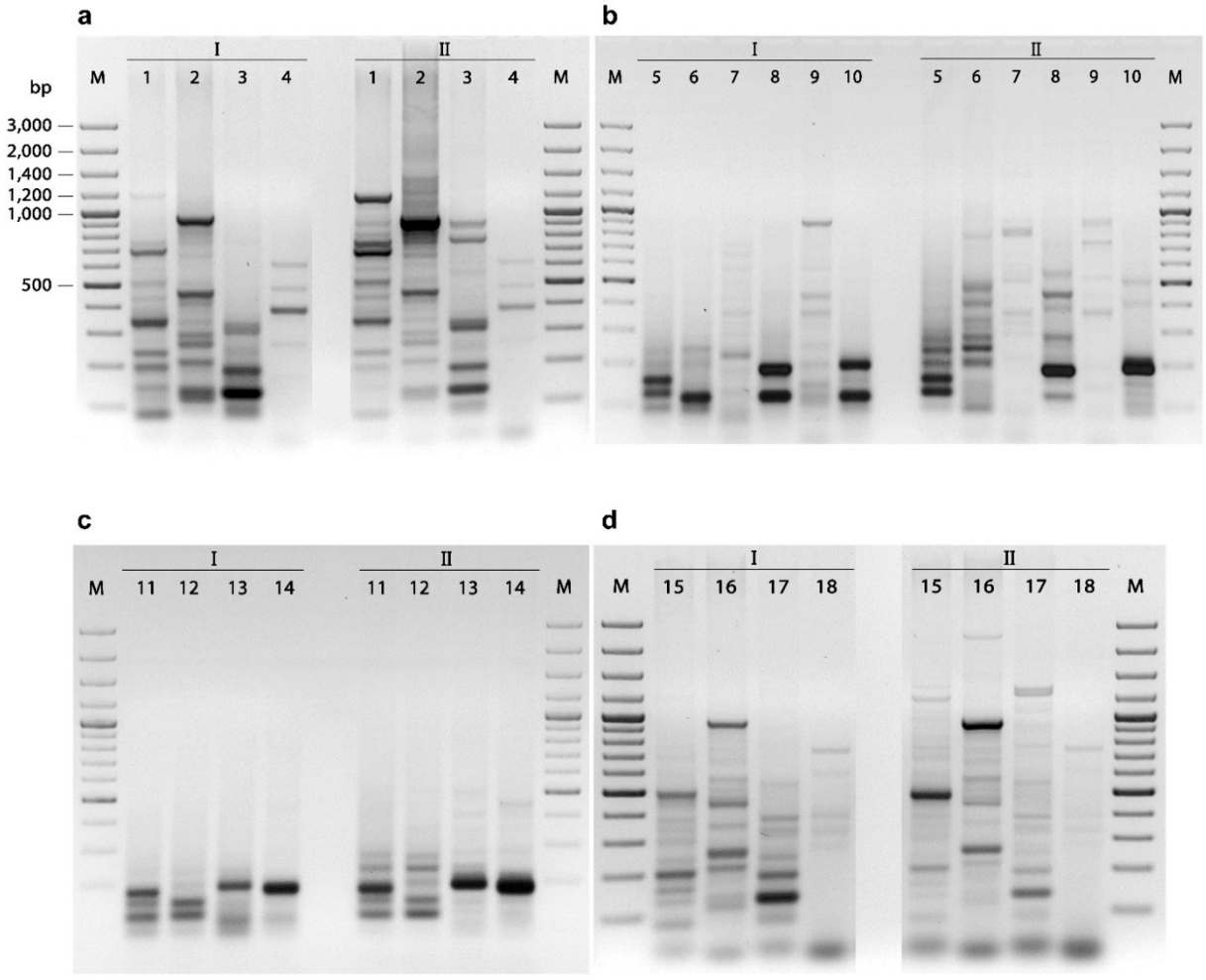
For ISRAP, genomic DNA of *L. decidua* was digested with the cost-effective restriction enzyme *Eco*RI (G/AATTC). *Eco*RI adapters were hybridized from the corresponding oligonucleotides (Table 3) and then ligated to the *Eco*RI-cleaved DNA fragments. These fragments were further

digested with the more frequently cutting restriction endonuclease *Mse*I (T/TAA), to avoid an excess of amplicons originating exclusively from *Eco*RI adapter sequences. Hence, the second DNA digestion supports the generation of PCR products originating from genomic regions between a SINE and an *Eco*RI-specific cleavage site. The *Eco*RI adapter primer was combined with a different number of SINE-derived primers in a single PCR (Figure 4).

Similar to the AFLP fingerprint technique (Vos *et al.*, 1995; Huang and Sun, 1999), the number of PCR fragments can be adjusted by a variable number of arbitrary 'selective nucleotides' at the 3' end of the *Eco*RI adapter primer. Each selective nucleotide reduces the number of amplicons by 25 %, since primer binding starts at the 3' end, which is crucial for elongation by the Taq DNA polymerase.

The combination of *Eco*RI adapter primer containing the three selective nucleotides 'GAC' (*Eco*RI-adap-GAC) and a single SINE-derived primer each (Figure 4a) generates strong bands in the range of ~ 100 bp to ~ 1,400 bp. As already observed for ISAP, the usage of 5' GC-rich primer extensions in combination with the respective PCR program (ISAP/ISRAP PCR) improves some of the banding patterns (Figure 4a, 1 and 3). An increased number of SINE-derived primers leads to a proportionally reduction of bands (Figure 4a-c). Using two SINE-derived primers together with the *Eco*RI-adap-GAC primer, bands larger than 1,000 bp are absent and the bands ranging between 500 bp and 1000 bp are relatively weak (Figure 4b). While the usage of two SINE-derived primers might be still appropriate, the application of three SINE-derived primers generates only strong bands below 300 bp (Figure 4c) and might not be sufficient for a genotype comparison.

The application of an *Eco*RI adapter primer containing two selective nucleotides (*Eco*RI-adap-GA) in combination with one SINE-derived primer produced a similar number of bands compared to the *Eco*RI-adap-GAC primer, but mostly of less intensity (Figure 4a, d).



a	ISAP primer	selective bases
1	LdS-II_*for*	GAC
2	LdS-II_*rev*	GAC
3	LdS-III_*for*	GAC
4	LdS-III_*rev*	GAC

b	ISAP primer		selective bases
5	LdS-II_*for*	LdS-II_*rev*	GAC
6	LdS-II_*for*	LdS-III_*for*	GAC
7	LdS-II_*for*	LdS-III_*rev*	GAC
8	LdS-II_*rev*	LdS-III_*for*	GAC
9	LdS-II_*rev*	LdS-III_*rev*	GAC
10	LdS-III_*for*	LdS-III_*rev*	GAC

c	ISAP primer			selective bases
11	LdS-II_*for*	LdS-II_*rev*	LdS-III_*for*	GAC
12	LdS-II_*for*	LdS-II_*rev*	LdS-III_*rev*	GAC
13	LdS-II_*for*	LdS-III_*for*	LdS-III_*rev*	GAC
14	LdS-II_*rev*	LdS-III_*for*	LdS-III_*rev*	GAC

d	ISAP primer	selective bases
15	LdS-II_*for*	GA
16	LdS-II_*rev*	GA
17	LdS-III_*for*	GA
18	LdS-III_*rev*	GA

Figure 4. Examination of ISRAP conditions suitable for genotype comparison. A specimen of *L. decidua*, growing at the Forest Botanical Garden of Tharandt, was examined using *Eco*RI adapter primers with three (a-c) or two (d) selective nucleotides combined with either one (a, d), two (b), or three SINE-derived primers (c). The banding patterns of standard PCR (I) were compared with those of the ISRAP PCR (II). The annealing temperature for all reaction was 50 °C. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

Compared to ISAPs (Figure 1b), the number of bands generated by ISRAP is generally decreased, while the band intensity is improved as intended (Figure 4). The combination of *Eco*RI-adap-GAC with a single SINE-derived primer (Figure 4a) was selected for a genotype comparison to examine the polymorphism count (Figure 5).

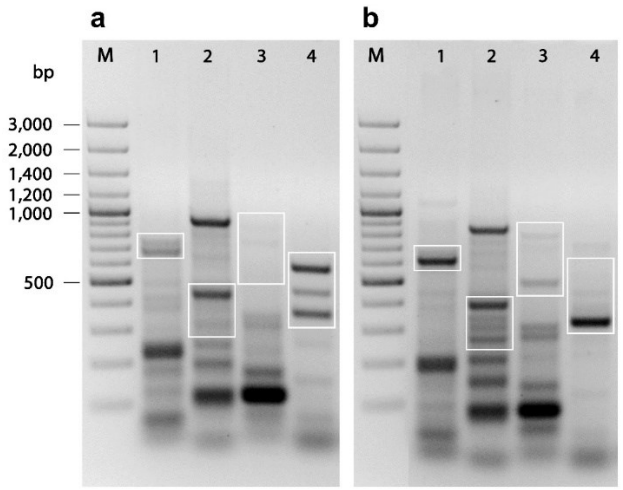


Figure 5. ISRAP profiles of two *L. decidua* genotypes with varying SINE-derived primers. The *L. decidua* genotype 'Tharandt' (a) was compared with *L. decidua* 'breeding number 91' obtained from the Saxony State Forestry Service (b). The *Eco*RI-adap-GAC primer was combined with SINE-derived primers using standard PCR: 1 - LdS-II_*for*, 2 - LdS-II_*rev*, 3 - LdS-III_*for*, 4 - LdS-III_*rev*. The annealing temperature for all reaction was 50 °C. Variable regions within the banding patterns are framed in white. The size marker (M) 'GeneRuler™ 100 bp Plus DNA Ladder' was used.

The *L. decidua* genotypes 'Tharandt' and 'breeding number 91' (b-no.91) show different banding patterns for all SINE-derived primers investigated. Significant polymorphisms are caused by the SINE-derived primers LdS-II_*for* and LdS-III_*rev* (Figure 5, 1 and 4). The variations of the ISRAP profiles generated by LdS-II_*rev* and LdS-III_*for* are not necessarily polymorphisms. The missing bands in the *L. decidua* 'Tharandt' banding pattern might also result from quality differences of genomic DNAs between both genotypes.

As the resolution of agarose gel electrophoresis is not sufficient, numerous ISRAP profiles of different primer combinations would have to be combined to achieve the discrimination between highly similar genotypes. Instead, the information content of a single ISRAP assay (band count vs. peak count) could be clearly increased by an amplicon separation using the capillary electrophoresis-based fragment length analysis (FLA) service.

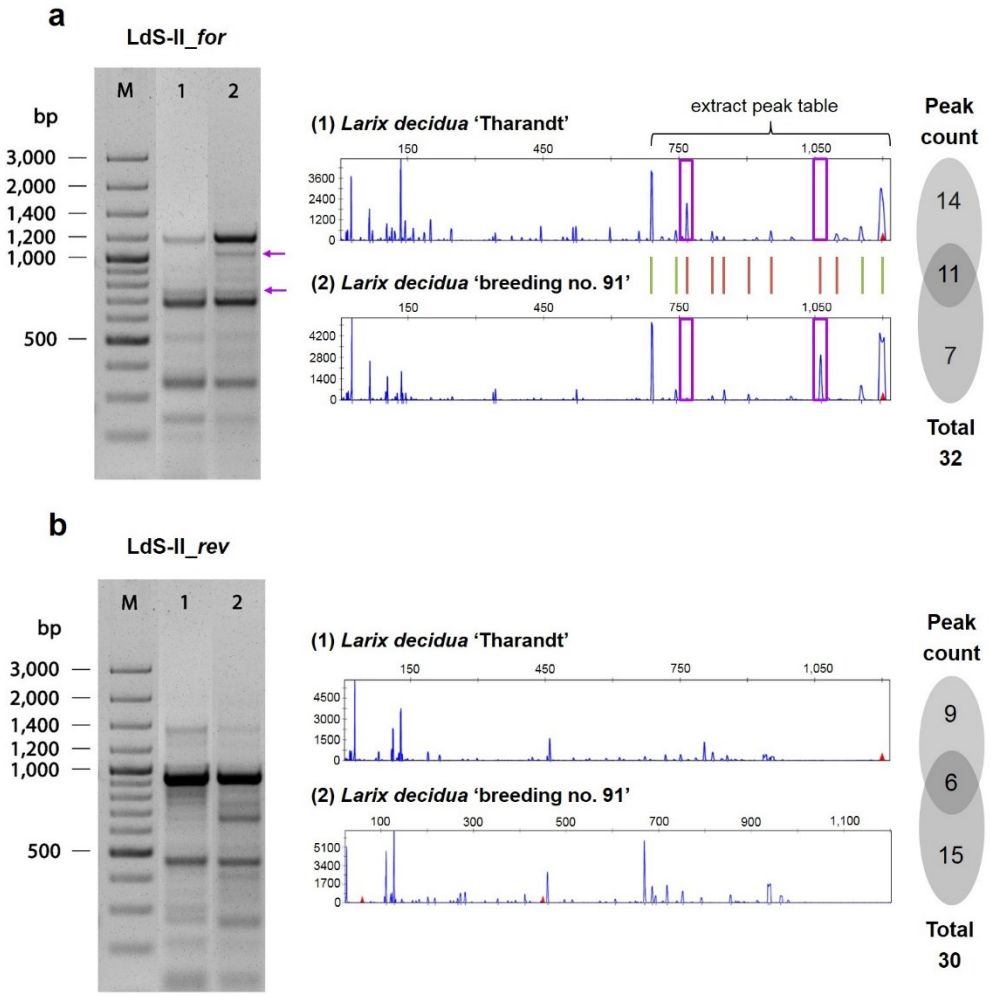
Figure 6 compares the sensitivity of both separation methods using the combinations of *Eco*RI-adap-GAC with LdS-II_*for* and LdS-II_*rev*, respectively. The number of peaks in the resulting electropherogram is defined by the range of the size standard LIZ-1200 (20 bp - 1,200 bp) and the signal intensity threshold of 400 relative fluorescence units (rfu) to distinguish between peaks and background noise (Figure 6; Supplementary chapter, Table S1, Figure S1 - S2).

The ISRAP-based comparison of the genotypes Tharandt and b-no.91 using the SINE-derived primer LdS-II_*for* resulted in the identification of 32 peak size classes (Figure 6a). Eleven peaks are shared by both genotypes and 21 of the total peak count are polymorphic with 14 Tharandt-specific and seven b-no.91-specific peaks. An exemplarily supplied extract from the respective peak table (Table S1) shows the size classes 22 to 32 and demonstrates that common peaks vary in signal intensity (height) and to a defined variance of 4 bp in size (peak area).

The two most prominent differences between both genotype profiles are located at 767 bp and 1,063 bp (Figure 6a, peak table, framed in purple), as these polymorphic peaks show high signal intensity (> 2,000 rfu). The length polymorphism at 1,063 bp (size 29) can be retrieved as a respective band for b-no.91 on the agarose gel, while other peak classes cannot be directly compared with the situation on the agarose gel (Figure 6a, arrows). This banding pattern is also presented in correlation with the determination of suitable ISRAP conditions (Figure 4a, II, lane1).

Using the combination of *Eco*RI-adap-GAC primer and the SINE-derived primer LdS-II_*rev* 30 peak size classes were defined, which are composed of six common peaks and 24 polymorphisms (Figure 6b). The majority of polymorphic peaks and bands, respectively, is located between 600 bp and 900 bp.

These initial experiments demonstrate the potential of the combined ISAP and AFLP primer application, together with the more sensitive visualization of amplicons achieved by FLA, to increase the possibility of polymorphism detection.



extract peak table

peak class	LdS-II_*for*	
	Tharandt	b-no.91
Size 22	**690.07**	**690.44**
Height 22	3913	4918
Peak Area 22	53683.0	75266.0
Size 23	**742.82**	**742.5**
Height 23	636	712
Peak Area 23	16464.0	19198.0
Size 24	**767.23**	
Height 24	2185	
Peak Area 24	53558.0	
Size 25	**822.91**	
Height 25	585	
Peak Area 25	12041.0	
Size 26		**849.57**
Height 26		741
Peak Area 26		15946.0
Size 27		**903.01**
Height 27		426
Peak Area 27		8528.0
Size 28	**952.73**	
Height 28	624	
Peak Area 28	12719.0	
Size 29		**1062.53**
Height 29		3028
Peak Area 29		63532.0
Size 30	**1097.72**	
Height 30	428	
Peak Area 30	10182.0	
Size 31	**1152.29**	**1151.71**
Height 31	842	1011
Peak Area 31	18146.0	22233.0
Size 32	**1196.26**	**1192.96**
Height 32	3026	4411
Peak Area 32	89576.0	97111.0

Figure 6. Resolution of different gel electrophoreses with regard to ISRAP polymorphism density. The *L. decidua* genotypes Tharandt (lane 1) and breeding number 91 (lane 2) were compared with ISRAP using the LdS-II_*for* (a) and the LdS-II_*rev* (b) primer in combination with the *Eco*RI-adap-GAC primer. The resulting amplicons were separated by agarose gel (left, M - size marker 'GeneRuler™ 100 bp Plus DNA Ladder') and capillary electrophoresis (right). Exemplary for ISRAP using LdS-II_*for* (a), an extract of the respective peak table illustrates shared peaks (green) and polymorphisms (red).The two most distinct polymorphisms in (a) are framed in the peak table (purple) and the corresponding positions on the agarose gel are indicated by arrows. The information content of the fragment length analysis (peak count) is represented as Venn diagram showing the number of shared and genotype-specific peaks of the respective peak profiles (electropherograms).

Discussion

Compared to the parent species, hybrid larches (*Larix* × *eurolepis*) display superior properties due to 'heterosis', or also called 'hybrid virgor' (Marchal *et al.*, 2017). However, the progeny of European (*Larix decidua*) and Japanese larch (*Larix kaempferi*) crosses contains less hybrid genotypes (Lee, 2003), which have to be identified by marker-assisted selection.

The earliest attempts of characterizing the progeny of larch crosses to select planting stock of pure hybrids used isoenzyme markers (Hacker and Bergmann, 1991; Ennos and Qian, 1994). Subsequently, also a morphological differentiation of seedlings was introduced (Pâques *et al.*, 2006), which is difficult and often not unambiguously. The development of molecular markers was supposed to provide fast, cost-effective and reliable results.

Initially, the hybrid fraction of *Larix* crosses was estimated with RAPD markers (Scheepers *et al.*, 2000). Also, the combination of maternally inherited mtDNA markers and paternally inherited cpDNA markers was used for the identification of hybrid larch genotypes (Acheré *et al.*, 2004) and to measure the rate of spontaneous hybridization (Meirmans *et al.*, 2014). Gros-Louis *et al.* (2005) used mtDNA, cpDNA and nuclear gene sequences to develop species-specific markers and combined them with RAPDs.

Furthermore, SSR markers were developed for Japanese (Isoda and Watanabe, 2006; Yang *et al.*, 2011; Li *et al.*, 2014; Chen *et al.*, 2015) and later also for European Larch (Wagner *et al.*, 2012; Nardin *et al.*, 2015; Gramazio *et al.*, 2018).

In order to distinguish hybrid seedlings from those corresponding to one of the parental genomes, it was intended to establish the ISAP marker system for European and Japanese larch, respectively, as a basis for the development of combined *L. decidua* and *L. kaempferi* ISAP primer combinations for the targeted identification of *L.* × *eurolepis* genotypes in crossbred offsprings.

The *Larix decidua* ISAP profiles show an insufficient resolution for genotyping

Due to weak bands and a low polymorphism count (Figure 2b), the application of the ISAP marker system is less suitable for genotyping of *Larix decidua* accessions.

Little is known about SINEs in conifers. Despite the description of single SINE families (Au, PinS-I) in a few gymnosperm species like *Cycas revoluta*, *Ginkgo biloba*, *Chamaecyparis pisifera*, *Ephedra ciliata*, *Picea glauca*, *Picea sitchensis*, *Pinus taeda* (Fawcett *et al.*, 2006; Wenke *et al.*, 2011; Yagi *et al.*, 2011), SINEs have not been comprehensively studied, mainly due to the extreme large genome sizes of ~ 12 - 30 Gb (Kuzmin *et al.*, 2019).

Currently, genome draft assemblies are available for white spruce (*Picea glauca*) (Birol *et al.*, 2013), Norway spruce (*Picea abies*) (Nystedt *et al.*, 2013), loblolly pine (*Pinus taeda*) (Neale *et al.*, 2014; Wegrzyn *et al.*, 2014; Zimin *et al.*, 2014), and Siberian larch (*Larix sibirica*) (Kuzmin *et al.*, 2019). As a fundamental difference, gymnosperm genomes were not frequently reshaped by partial or whole genome duplications (WGDs) like observed for angiosperms. Consequently, their genomes show a highly stable macrostructure and were predominantly enlarged by TE proliferation, while tandem repeats have less contributed (reviewed in Wang and Ran, 2014).

Due to a genome size of ~ 13 Gb (Zonneveld, 2012) and the absence of long sequencing reads, the assembly of the *Larix decidua* genome sequences was not feasible. Thus, the small size of SINEs was utilized for their identification based on the Illumina sequencing reads of two different insert size libraries (Table 1, insert sizes of 300 bp and 470 bp). Sufficient read coverage provided, the 83 bp - 352 bp long SINEs (Deragon and Zhang, 2006; Wenke *et al.*, 2011) are statistically located within the library fragments enabling their detection by paired-end sequencing (Figure 1). Since the concatenated 101 bp forward and reverse reads were screened with the *SINE-Finder*, the complete SINE consensus sequences had to be detected by *BLAST* searches in the read database. With an approximately 5-fold genome coverage (Table 1), the most abundant SINE families have most likely been identified. Due to the absence of full-length copies, the SINE families could not be characterized concerning the copy numbers and similarity. The number of *SINE-Finder* output sequences only gives rough indications, which SINE families might be suitable for the ISAP primer design. A low sequence count does not imply low abundance like observed for *C. japonica* (Chapter 2.1, Table 3 and 4). The *SINE-Finder*

cluster may contain only few sequences (Chapter 2.1, Table 3, cluster 10, 24, 27), but the abundance of the respective SINE family might be in a range suitable for ISAP primer design (Chapter 2.1, Table 4, TheaS-VII, -X, and -XIII) together with appropriate similarity (Chapter 2.1, Table 4, TheaS-VII). Hence, ISAP primers based on the remaining *L. decidua* SINE families PinS-I, PinS-IV, and PinS-V should be additionally tested to complete the ISAP establishment procedure.

Inter-SINE-restriction site amplified polymorphism (ISRAP) – Development of a novel marker technique

Due to the high suitability of the AFLP technique for the analysis of hybrid genome compositions (Burdon and Wilcox, 2011), the number of fragment length polymorphisms based on SINE distribution was intended to be increased by the combination of ISAP with AFLP primers (Vos *et al.*, 1995). In contrast to AFLP, the *Mse*I adapter ligation was omitted. Instead, amplicons were created based on the *Eco*RI adapters and the SINEs located on the EcoRI / Msel fragments.

The comparison of the *L. decidua* genotypes ‘Tharandt’ and ‘breeding no. 91’ with ISRAP assays using four ISAP primers, respectively, showed polymorphic bands on the agarose gel, particularly evident for PinS-III*_rev* (Figure 5, lane 4). Moreover, the sensitivity for the detection of PCR amplicons could be increased using the fragment length analysis (FLA) service (Eurofins Genomics, Ebersberg, Germany). Initial experiments revealed 21 polymorphisms for the SINE-derived primer LdS-II*_for* and 24 polymorphic peaks using LdS-II*_rev* (Figure 6).

However, to ensure robust results, the reproducibility of the peak profiles has to be verified in several repetitions with regard to the stability of the peak pattern and to measure fluctuations of the peak intensity. If necessary, the signal intensity threshold has to be increased to guarantee a clear separation between signal peaks and background noise.

The results of agarose and capillary gel electrophoresis cannot be directly compared, as exemplified by the comparative amplicon visualization in Figure 6a. While the agarose gel shows all PCR products, the FLA only displays amplicons resulting from at least one labeled *Eco*RI-adap-GAC primer. Amplicons derived from an inter-SINE region might be rare, but cannot be fully excluded.

The FLA provides a clearly increased amplicon resolution and reduces the manual effort. As amplicons from different ISRAPs can be pooled using four different fluorescent dyes, no additional costs arise (currently 2.60 Euro per genotype). However, if ISRAPs provide significantly more polymorphisms than ISAPs had to be evaluated by simultaneous FLAs.

Adjusting screws and possible improvements for ISRAP assays

Two major modifications are possible to regulate the number of amplicons:

(I)

The combination of the *Eco*RI adapter primer with a single ISAP primer provides the most suitable results for genotype comparison. The application of two or even three SINE-derived primers together with the *Eco*RI adapter primer resulted in a preferred synthesis of smaller amplicons and hence, in a reduced band count (Figure 4a-c). The usage of *Eco*RI adapter primer with two and three selective bases, respectively, generated altered fingerprints of similar band count without affecting the number of polymorphisms on the agarose gel (Figure 4a, d). For AFLP analyses usually three selective nucleotides are sufficient to reduce the number of PCR products (Vos *et al.*, 1995). For ISRAP, the usage of both, two and three selective nucleotides, respectively, provided results appropriate for evaluation (Figure 4a, d).

(II)

The polymorphism rate might be mainly influenced by the respective SINE-derived primer. Hence, LdS primers of the remaining PinS families should be examined and compared with those of PinS-II (LdS-II_*for* / LdS-II_*rev*). Especially SINE families with evolutionarily young copies enable polymorphic bands and have to be selected for the ISRAP approach. Furthermore, the combination of the *Eco*RI-adapter primers with two SINE-derived primers, for example *L. decidua* and *L. kaempferi*-specific each, might additionally increase the differentiation capacity for hybrid seedlings. However, Japanese and European larch do not show substantial interspecific variations (Semerikov *et al.*, 1999; Acheré *et al.*, 2004), indicating that *L. decidua* -derived primers might be sufficient for the differentiation between hybrid larch and parental genotypes.

The standard AFLP enzyme combination is composed of *Eco*RI (G/AATTC) for the initial DNA cleavage and the more frequently cutting enzyme *Mse*I (T/TAA) for the second digestion. More cost-effective restriction endonucleases with a similar cleavage frequency ratio might be tested for further reduction of costs like the replacement of *Mse*I by *Bsu*RI (GG/CC, not methylation-sensitive). Furthermore, the nucleotide composition of the recognition site might also affect the amount of polymorphisms, for example by base-specific mutation rates or depending on the genome-wide distribution of the GC content.

Alternative marker approaches, derived from the AFLP principle, were developed, for example in combination with SSRs (Witsenboer *et al.*, 1997) or non-autonomous transposons (Park *et al.*, 2003). Retrotransposons were first combined with the AFLP technique using the Ty1-*copia*-like *BARE*-1 family (Manninen and Schulman, 1993) in barley, designated sequence-specific amplification polymorphism (S-SAP) (Waugh *et al.*, 1997). The *Mse*I/*Pst*I-digested genomic DNA of two barley genotypes was amplified using one of the respective adapter primers, containing up to three selective nucleotides, combined with radiolabeled primers originating from the 5′end of the *BARE*-1 LTR sequence. Compared to AFLP assays in barley (Powell *et al.*, 1997), the total number of fragments was lower, while the polymorphism count was in a similar range with an average of eight vs. eleven polymorphisms, respectively.

Hence, the magnitude of the polymorphism count achieved by S-SAP and ISRAP (21 - 24 polymorphisms) might be in a similar range, although the results are not directly comparable due to the application in different species, differing restriction endonuclease combinations and fragment detection methods. However, similar to SINEs, the *BARE*-1 LTRs show a dispersed distribution throughout the genome, less frequently occurring in centromere regions and often locally clustered or nested (Waugh *et al.*, 1997), which might explain the comparable results.

S-SAPs were also tested with other retrotransposons and restriction enzymes (Leigh *et al.*, 2003) and applied in a broad range of plants, for example wheat (Queen *et al.*, 2004), apple (Venturi *et al.*, 2006), lettuce (Syed *et al.*, 2006), pea (Jing *et al.*, 2005) as well as pepper and tomato (Tam *et al.*, 2005).

Furthermore, molecular markers were established using the LTR-derived primers of the *BARE*-1 family solely (inter-retrotransposon amplified polymorphism, IRAP) and in combination with microsatellite loci (retrotransposon-microsatellite amplified polymorphism, REMAP) (Kalendar *et al.*, 1999; Campbell *et al.*, 2011).

However, the retrotransposon-derived molecular markers were less frequently used as the most prominent techniques (SSR, SNP, DArT) mostly cover the required spectrum of marker applications in plant breeding (Burdon and Wilcox, 2011). Perspectively, more efficient SNP genotyping techniques, for example genome-wide association studies (GWAS) (Zheng *et al.*, 2017; Kim *et al.*, 2018), will replace SSRs in parentage analyses, although they are still unaffordable for routine sorting (Burdon and Wilcox, 2011; De La Torre *et al.*, 2014).

References

Acheré, V., Faivre Rampant, P., Pâques, L.E. and Prat, D. (2004) Chloroplast and mitochondrial molecular tests identify European × Japanese larch hybrids. *Theor. Appl. Genet.*, **108**, 1643–1649.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Birol, I., Raymond, A., Jackman, S.D., et al. (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**, 1492–1497.

Burdon, R. and Wilcox, P. (2011) Integration of molecular markers in breeding. In C. Plomion, J. Bousquet, and C. Kole, eds. *Genetics, genomics and breeding of conifers.* New York: CRC Press and Edenbridge Science Publishers, pp. 276–322.

Campbell, B.C., LeMare, S., Piperidis, G. and Godwin, I.D. (2011) IRAP, a retrotransposon-based marker system for the detection of somaclonal variation in barley. *Mol. Breed.*, **27**, 193–206.

Chen, X.-B., Xie, Y.-H. and Sun, X.-M. (2015) Development and characterization of polymorphic genic-SSR markers in *Larix kaempferi*. *Molecules*, **20**, 6060–6067.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Eko, P.M., Larsson-Stern, M. and Albrektson, A. (2004) Growth and yield of hybrid larch (*Larix* × *eurolepis* A. Henry) in Southern Sweden. *Scand J Forest Res*, **19**, 320–328.

Ennos, R.A. and Qian, T. (1994) Monitoring the output of a hybrid larch seed orchard using isozyme markers. *For. An Int. J. For. Res.*, **67**, 63–74.

Fawcett, J.A., Kawahara, T., Watanabe, H. and Yasui, Y. (2006) A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant Mol. Biol.*, **61**, 505–514.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Gramazio, P., Plesa, I.M., Truta, A.M., Sestras, A.F. and Vilanova, S. (2018) Highly informative SSR genotyping reveals large genetic diversity and limited differentiation in European larch (*Larix decidua*) populations from Romania. *Turk J Agric*, **42**, 165–175.

Gros-Louis, M.-C., Bousquet, J., Pâques, L.E. and Isabel, N. (2005) Species-diagnostic markers in

Larix spp. based on RAPDs and nuclear, cpDNA, and mtDNA gene sequences, and their phylogenetic implications. *Tree Genet. Genomes*, **1**, 50–63.

Hacker, M. and Bergmann, F. (1991) The proportion of hybrids in seed from a seed orchard composed of two larch species (*L. europaea* and *L. leptolepis*). *Ann. For. Sci.*, **48**, 631–640.

Harrison, A.J., Hoffmann, D., Kannenberg, N., Lelu, M.A., Verger, M., Le Pichon, C. and Bourlon, V. (2002) Developments in hybrid larch (*Larix* × *eurolepis* Henry) vegetative propagation in North Western Europe. In L.E. Pâques, ed. *Improvement of larch* (Larix sp.) *for better growth, stem form and wood quality.* Proceedings of Larix 2002 IUFRO Symposium, France. INRA, Oliver Cedex, France, pp. 250–265

Henry, A. and Flood, M.G. (1919) The history of the dunkeld hybrid larch, *Larix* × *eurolepis*, with notes on other hybrid conifers. *Proc. R. Ir. Acad. B.*, **35**, 55–66.

Huang, J. and Sun, M. (1999) A modified AFLP with fluorescence-labelled primers and automated DNA sequencer detection for efficient fingerprinting analysis in plants. *Biotechnol. Tech.*, **13**, 277–278.

Isoda, K. and Watanabe, A. (2006) Isolation and characterization of microsatellite loci from *Larix kaempferi*. *Mol. Ecol. Notes*, **6**, 664–666.

Jing, R., Knox, M.R., Lee, J.M., Vershinin, A. V, Ambrose, M., Ellis, T.H.N. and Flavell, A.J. (2005) Insertional polymorphism and antiquity of PDR1 retrotransposon insertions in *Pisum* species. *Genetics*, **171**, 741–752.

Kalendar, R., Grob, T., Regina, M., Suoniemi, A. and Schulman, A. (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theor. Appl. Genet.*, **98**, 704–711.

Kim, B., Udvardi, M.K., Zhang, W., et al. (2018) GWASpro: a high-performance genome-wide association analysis server. *Bioinformatics*, bty989, doi: 10.1093/bioinformatics/bty989.

Klimaszewska, K. (1989) Recovery of somatic embryos and plantlets from protoplast cultures of *Larix* × *eurolepis*. *Plant Cell Rep.*, **8**, 440–444.

Kraft, A. and Kadolsky, M. (2018) Hybrid larch (*Larix* × *eurolepis* Henry). In S. M. Jain and P. Gupta, eds. *Stepwise protocols for somatic embryogenesis of important woody plants, volume II.* Springer International Publishing, pp. 149–158.

Kuzmin, D.A., Feranchuk, S.I., Sharov, V.V., Cybin, A.N., Makolov, S. V., Putintseva, Y.A., Oreshkova, N.V. and Krutovsky, K.V. (2019) Stepwise large genome assembly approach: a

case of Siberian larch (*Larix sibirica* Ledeb.). *BMC Bioinformatics*, **20**, 37.

La Torre, A.R. De, Birol, I., Bousquet, J., et al. (2014) Insights into conifer giga-genomes. *Plant Physiol.*, **166**, 1724–1732.

Larsson-Stern, M. (2003) *Aspects of hybrid larch* (Larix × eurolepis *Henry*) *as a potential tree species in southern Swedish forestry.* Alnarp: Swedish University of Agricultural Sciences.

Lee, S.J. (2003) Breeding hybrid larch in Britain. In *Forestry commission information note 52.* Edinburgh: Forestry Commission, pp. 1–4.

Leigh, F., Kalendar, R., Lea, V., Lee, D., Donini, P. and Schulman, A.H. (2003) Comparison of the utility of barley retrotransposon families for genetic analysis by molecular marker techniques. *Mol. Genet. Genomics*, **269**, 464–474.

Lelu-Walter, M.-A. and Pâques, L.E. (2009) Simplified and improved somatic embryogenesis of hybrid larches (*Larix* × *eurolepis* and *Larix* × *marschlinsii*). Perspectives for breeding. *Ann. For. Sci.*, **66**, 104p1–104p10.

Lelu, M.A., Bastien, C., Klimaszewska, K., Ward, C. and Charest, P.J. (1994) An improved method for somatic plantlet production in hybrid larch (*Larix* × *leptoeuropaea*): part 1. Somatic embryo maturation. *Plant Cell. Tissue Organ Cult.*, **36**, 107–115.

Li, W., Han, S., Qi, L. and Zhang, S. (2014) Transcriptome resources and genome-wide marker development for Japanese larch (*Larix kaempferi*). *Front Agr Sci Eng*, **1**, 77–84.

Manninen, I. and Schulman, A.H. (1993) BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). *Plant Mol. Biol.*, **22**, 829–846.

Marchal, A., Muñoz, F., Millier, F., Sánchez, L. and Pâques, L.E. (2017) Hybrid larch heterosis: for which traits and under which genetic control? *Tree Genet. Genomes*, **13**, 92.

Matyssek, R. and Schulze, E.-D. (1987) Heterosis in hybrid larch (*Larix decidua* × *leptolepis*). *Trees*, **1**, 219–224.

Meirmans, P.G., Gros-Louis, M.-C., Lamothe, M., Perron, M., Bousquet, J. and Isabel, N. (2014) Rates of spontaneous hybridization and hybrid recruitment in co-existing exotic and native mature larch populations. *Tree Genet. Genomes*, **10**, 965–975.

Nardin, M., Musch, B., Rousselle, Y., et al. (2015) Genetic differentiation of European larch along an altitudinal gradient in the French Alps. *Ann. For. Sci.*, **72**, 517–527.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., et al. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.*, **15**, R59.

Nystedt, B., Street, N.R., Wetterbom, A., et al. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579.

Pâques, L.E. (2009) Growth rhythm parameters as components of hybrid vigour in young seedlings of hybrid larch (*Larix decidua* × *L. kaempferi*). *Silvae Genet.*, **58**, 42–53.

Pâques, L.E., Foffová, E., Heinze, B., Lelu-Walter, M.-A., Liesebach, M. and Philippe, G. (2013) Larches (*Larix* sp.). In L.E. Pâques, ed. *Forest tree breeding in Europe: current state-of-the-art and perspectives.* Dordrecht: Springer Netherlands, pp. 13–122.

Pâques, L.E., Philippe, G. and Prat, D. (2006) Identification of European and Japanese larch and their interspecific hybrid with morphological markers: application to young seedlings. *Silvae Genet.*, **55**, 123–134.

Park, K.C., Kim, N.H., Cho, Y.S., Kang, K.H., Lee, J.K. and Kim, N.-S. (2003) Genetic variations of AA genome *Oryza* species measured by MITE-AFLP. *Theor. Appl. Genet.*, **107**, 203–209.

Perks, M., Harrison, A., McKay, H. and Morgan, J. (2006) An update on nursery propagation and establishment best practice for larch in Britain. In *Forestry Commission information note 80.* Edinburgh: Forestry Commission, pp. 1–6.

Powell, W., Thomas, W.T.B., Baird, E., Lawrence, P., Booth, A., Harrower, B., McNicol, J.W. and Waugh, R. (1997) Analysis of quantitative traits in barley by the use of amplified fragment length polymorphisms. *Heredity (Edinb).*, **79**, 48–59.

Queen, R.A., Gribbon, B.M., James, C., Jack, P. and Flavell, A.J. (2004) Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat. *Mol. Genet. Genomics*, **271**, 91–97.

Scheepers, D., Eloy, M.-C. and Briquet, M. (2000) Identification of larch species (*Larix decidua*, *Larix kaempferi* and *Larix* × *eurolepis*) and estimation of hybrid fraction in seed lots by RAPD fingerprints. *Theor. Appl. Genet.*, **100**, 71–74.

Semerikov, V.L., Semerikov, L.F. and Lascoux, M. (1999) Intra- and interspecific allozyme variability in Eurasian *Larix* Mill. species. *Heredity*, **82**, 193.

Syed, N.H., Sørensen, A.P., Antonise, R., Wiel, C. van de, Linden, C.G. van der, `t Westende, W. van, Hooftman, D.A.P., Nijs, H.C.M. den and Flavell, A.J. (2006) A detailed linkage map of lettuce based on SSAP, AFLP and NBS markers. *Theor. Appl. Genet.*, **112**, 517–527.

Tam, S.M., Mhiri, C., Vogelaar, A., Kerkveld, M., Pearce, S.R. and Grandbastien, M.-A. (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. *Theor. Appl. Genet.*, **110**, 819–831.

Venturi, S., Dondini, L., Donini, P. and Sansavini, S. (2006) Retrotransposon characterisation and fingerprinting of apple clones by S-SAP markers. *Theor. Appl. Genet.*, **112**, 440–444.

Vos, P., Hogers, R., Bleeker, M., et al. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.*, **23**, 4407–4414.

Wagner, S., Gerber, S. and Petit, R.J. (2012) Two highly informative dinucleotide SSR multiplexes for the conifer *Larix decidua* (European larch). *Mol. Ecol. Resour.*, **12**, 717–725.

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

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.B.T. and Powell, W. (1997) Genetic distribution of Bare–1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol. Gen. Genet.*, **253**, 687–694.

Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., et al. (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, **196**, 891–909.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Witsenboer, H., Michelmore, R.W. and Vogel, J. (1997) Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca* spp.). *Genome*, **40**, 923–936.

Yagi, E., Akita, T. and Kawahara, T. (2011) A novel Au SINE sequence found in a gymnosperm. *Genes Genet. Syst.*, **86**, 19–25.

Yang, X., Sun, X. and Zhang, S. (2011) Short note: development of six EST-SSR markers in larch. *Silvae Genet.*, **60**, 161–163.

Zheng, M., Peng, C., Liu, H., et al. (2017) Genome-wide association study reveals candidate genes for control of plant height, branch initiation height and branch number in rapeseed (*Brassica napus* L.). *Front. Plant Sci.*, **8**, 1246.

Zimin, A., Stevens, K.A., Crepeau, M.W., et al. (2014) Sequencing and assembly of the 22 Gb loblolly pine genome. *Genetics*, **196**, 875–890.

Zonneveld, B.J.M. (2012) Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nord. J. Bot.*, **30**, 490–502.

Chapter 4

Summarizing Discussion - Application of SINE-based Marker Systems in Angiosperm and Gymnosperm Tree Species

The inter-SINE amplified polymorphism (ISAP) is a DNA fingerprinting technique based on the amplification of genomic DNA flanked by SINEs to detect amplicon length polymorphisms for genotyping. ISAPs provided reliable results for the discrimination of potato (*Solanum tuberosum*) cultivars (Seibt *et al.*, 2012). The SINE-based marker system was intended to be applied to angiosperm and gymnosperm tree species in order to examine relationships between natural populations, to perform parentage analyses and to analyze their potential for cross-species applications.

Nowadays, microsatellite markers are still the most frequently applied method of genotyping in plant breeding due to the ease of use and the high polymorphism rate (Jiang, 2013; Garrido-Cardenas *et al.*, 2018). However, the detection of informative simple sequence repeat (SSR) loci is more laborious and time-consuming (Jiang, 2013) than ISAP marker development, as SINEs are fast and easily detectable with the *SINE-Finder* tool (Wenke *et al.*, 2011), provided that a genome reference assembly is available.

In order to establish the ISAP method, initially two ISAP primers for at least three highly abundant and less diverse SINE families were designed for the species of interest. Different primer combinations were tested, and those creating amplicon length polymorphisms were applied for genotype comparisons. The ISAP establishment succeeded only for the angiosperm trees camellia (*Camellia japonica*) and European aspen (*Populus tremula*), but failed in the gymnosperm species European larch (*Larix decidua*), for which a novel marker system was developed based on ISAP and AFLP primers.

In this chapter the ISAP technique will be discussed according to the following topics:

4.1 Preconditions for successful ISAP applications

4.2 Reproducibility of ISAP profiles and potential sources of biased results

4.3 Future prospects

4.1 Preconditions for successful ISAP applications

The availability of at least partially assembled genome sequences is crucial for a fast SINE identification and the derivation of suitable ISAP primers (Table 1, *S. tuberosum*, *C. japonica* and *P. tremula*). The SINE identification based on short sequencing reads is also possible, but time-consuming and does not offer the opportunity for a profound SINE characterization (Table 1, *L. decidua*). Therefore, the selection of suitable SINE families for an ISAP primer design is vague, and a laborious examination of primer combinations derived from all SINE families identified would have to be conducted. The availablility of assembled genome sequences in public sequence databases depends on the economical importance of the species. Due to computational capacity, it used to be also dependent on the genome size. Hence, genome assemblies of the large gymnosperm genomes were successively provided with some time lag (Wang and Ran, 2014).

Table 1. Comparison of ISAP preconditions.

Species	Genome		SINE[f]			DNA
	assembly[a]	size (Mb)	families	abundance[h]	similarity [%]	extraction
Solanum tuberosum	yes	727[b]	9[g]	213 - 216[g]	77 - 87[g]	easy[i]
Camellia japonica	partially	2,300[c]	13	146 - 526	71 - 81	critical[j]
Populus tremula	yes	480[d]	7	33 - 1,174	75 - 84	critical[k]
Larix decidua	no	13,008[e]	6	N/A	N/A	critical[j]

a Availability of assembled genome sequences
b The Potato Genome Sequencing Consortium, 2011
c Huang *et al.*, 2013
d Lin *et al.*, 2018
e Zonneveld, 2012
f SINE families used for ISAP primer design, *S. tuberosum*: SolS-IIIa, SolS-IV (Seibt *et al.*, 2012)
g Wenke *et al.*, 2011
h Number of full-length SINE copies
i CTAB protocol (Saghai-Maroof *et al.*, 1984)
j Commercial kits
k SDS-based protocol (Verbylaite *et al.*, 2010)
N/A Not available

Typically, plant genomes do not harbor large SINE proportions. The SINE content in Poaceae genomes ranges between 0.005 % (*T. aestivum*) and 0.1 % (*O. sativa*) (Chapter 2.2, Figure 1, estimation based on full-length copies). In potato, SINEs occupy approximately 0.32 % of the genome (Seibt *et al.*, 2016) and the European aspen contains at least 0.19 % SINEs (Chapter 2.3, Figure 1, estimation based on full-length copies). However, the SINE proportion of the genome is presumably less decisive than expected. SINEs integrate randomly, however, they are not evenly distributed in the

genome and accumulate preferably in gene-rich, distal chromosome regions (Okada, 1991; Lenoir *et al.*, 2001; Wenke *et al.*, 2011). Due to the tendency to cluster (Jurka *et al.*, 2005; Seibt *et al.*, 2012), regions of high SINE density emerge resulting in smaller distances between adjacent SINE copies that can be amplified by PCR.

Highest importance for successful ISAP applications is associated with the SINE landscape of the respective genome. The total number of SINE families is less relevant, but at least two SINE families or subfamilies have to consist of approximately more than 200 highly similar full-length copies, as observed for potato (Table 1), to provide a sufficient number of polymorphic bands for genotype discrimination. Only the number of SINE full-length copies is considered, as frequently occurring related sequence regions among the SINE families often impede the correct assignment of the 5' truncated copies. The inter-SINE amplicon length polymorphisms might result from insertions, deletions, and genome rearrangements between adjacent SINEs as well as from mutations of the primer binding sites within the SINE copies.

In *C. japonica* and *P. tremula*, the similarity of the SINE families chosen for primer design does not exceed 81 % and 84 %, respectively (Table 1), analogue to the *S. tuberosum* SINE families SolS-IIIa and SolS-IV with 87 % and 77 % similarity, respectively (Wenke *et al.*, 2011; Seibt *et al.*, 2012). ISAP primers derived from these SolS families created the highest number of amplicons, thereby showing the highest density of polymorphisms (Seibt *et al.*, 2012). In *C. japonica* especially TheaS-II (81 % similarity and 526 full-length copies) is highly suitable for ISAP primer design. However, primers of TheaS-I, -II, and -IV (CjS-I*_for*, CjS-II*_for*, and CjS-IV*_rev*) equally contributed to the most informative banding patterns (Chapter 2.1, Table 4; Chapter 3.1, Figure 3). In *P. tremula* the two most abundant SINE families with corresponding ISAP primers are SaliS-I with 1,174 full-length copies and SaliS-IV.1 with 902 full-length copies (Chapter 2.3, Figure 1). Their similarities are comparatively low with 75 % and 76 %, respectively (Chapter 2.3, Table S5). Moreover, due to the shared 3' regions of the SaliS families (Chapter 3.2, Figure 1a), polymorphic ISAP profiles were even achieved using the *Populus tremula* SINE (PtS) ISAP primer PtS-I*_for* solely. Usually, the application of a single ISAP primer is less efficient: In *Camellia japonica* only the ISAP primer CjS-IV*_rev*

generated several bands without contribution of other SINE families showing sequence homologies (Chapter 3.1, Figure 3, 8/8).

The length of the SINE families has also a key influence on the efficiency of the ISAP primers. Based on the *SINE-Finder* output, PinS-VI might be one of the most abundant SINE families in the *L. decidua* genome and exhibits a short conserved length of 126 bp (Chapter 3.3, Figure 2a). As it is recommended to derive ISAP primers from the tRNA-unrelated 3' region (Wenke *et al.*, 2015), only 45 bp are left for placing the two outward-facing primer sequences. Due to the general guidelines for primer design like similar GC content above 50 %, an evenly distributed nucleotide content, avoidance of self-priming and formation of heterodimers, the *Larix decidua* SINE (LdS) ISAP primer LdS-VI_*rev* had to cover parts of the 5' region without containing highly conserved nucleotides of the RNA polymerase III promotor box B motif (Galli *et al.*, 1981). The same applies to CjS-II_*rev* derived from TheaS-II (Chapter 3.1, Figure 2a).

The SINE 3' region is also preferred for primer design as it increases the possibility to involve more SINE copies, in particular the 5' truncated SINEs. They originate from aborted reverse transcription, starting at the SINE 3' end (Luan *et al.*, 1993) and are usually as frequent as full-length copies (Myouga *et al.*, 2001; Lenoir *et al.*, 2001; Wenke *et al.*, 2011).

The extraction of pure, intact genomic DNA is a basic requirement for any type of molecular investigation. Initial ISAP experiments for tree species with *C. japonica* showed that the reproducibility of the banding pattern strongly depends on quality and quantity of the genomic DNA (Chapter 3.1, Figure 2d, 1 - 2). Despite the application of species-specifically adapted DNA extraction methods (Table 1) including subsequent purification as well as adapted DNA concentrations and stable PCR conditions, ISAP banding patterns with weak or missing bands still indicate the presence of inhibitory compounds (Chapter 3.1, Figure 2d).

Accordingly, it has been shown that DNA extracts from needles and leaves of mature trees sometimes contain high concentrations of secondary metabolites like polysaccharides, polyphenols, terpenes and tannins, which are hardly to remove (Shepherd *et al.*, 2002; Yoon *et al.*, 2017). Several specialized protocols have been developed and comparatively analyzed to enhance the success of DNA extraction

(Katterman and Shattuck, 1983; Ostrowska *et al.*, 1998; Tibbits *et al.*, 2006; Verbylaite *et al.*, 2010). However, yield and purity of the extracted DNA also depends on the tissue type and age and can vary among species of the same genus (Henry, 2001; Moreira and Oliveira, 2011).

Hence, as impurities cannot be completely avoided without immense efforts of time and costs, two PCR additives were supplied for ISAP analyses on tree species. Bovine serum albumin (BSA) is a standard ingredient of ISAP PCRs (Seibt *et al.*, 2012; Wenke *et al.*, 2015) and prevents interactions between the Taq DNA polymerase and secondary compounds (Kramer and Coen, 2006; Woide *et al.*, 2010; Farell and Alexandre, 2012). Betaine was added to each reaction, as it increases yield and specificity of PCR products based on facilitated strand separation (Frackman *et al.*, 1998).

Nevertheless, co-extracted contaminants substantially determine the success of genotyping based on ISAP, which is especially relevant for tree species (Table 1).

4.2 Reproducibility of ISAP profiles and potential sources of biased results

The reproducibility of ISAP fingerprints strongly depends on stable PCR and agarose gel electrophoresis conditions. The cooperation with the Saxony State Forestry Service (Pirna, Germany) and the group Molecular Physiology of Woody Plants of the Dresden University of Technology (Tharandt, Germany) revealed that the transfer of the ISAP method to other laboratories is associated with some difficulties. Reproducible results were achieved by coordinated experiment procedures and materials, like the same type of Taq DNA polymerase (DreamTaq™ Thermo Scientific, Waltham, USA) and LE agarose (Biozym Scientific GmbH, Oldendorf, Germany), usage of Eppendorf PCR cycler (Hamburg, Germany) with comparable heating and cooling rates and primer synthesis by Eurofins Genomics (Ebersberg, Germany).

Furthermore, the discrimination between highly similar (Chapter 3.1, Figure 4a) and identical (Chapter 3.1, Figure 4b) genotype profiles sometimes is hampered by inaccuracies resulting from each agarose gel electrophoresis that cannot be fully normalized with the fingerprint software *BioNumerics* (Applied Math, NV, Belgium). The banding patterns have to be compared and interpreted regarding the whole pattern range, which is not feasible with the applied software. Consequently, the size

assignment of the bands sometimes required manual corrections, which would be hardly to realize for high throughput applications.

The automated size assignment is also hampered by varying band intensities. The band visibility can be improved by editing the tone curve and band search settings like sensitivity can be regulated to guarantee standardized detection. However, the varying band intensities might have several reasons, some of which more obstructive for genotype comparisions than others.

(I)

The banding patterns of PCR-based multi-locus methods are regarded as dominant inherited markers. Comparisons of heterozygote and homozygote individuals might include varying intensity of some ISAP bands due to the effect of allele dosage (Nybom, 2004, 2014).

(II)

Band intensity differences may also result from non-specifically bound primers or diversified SINE copies, leading to irregular primer bindings for a specific locus.

(III)

Substantial biased ISAP results might predominantly arise in case of insufficient DNA purity, since parts of the banding patterns become faint or even undetectable. Thus, band information is missing and will be interpreted as nonexistent. Although some weak bands might be subsequently involved in the analysis by manual corrections, such processing steps are time-consuming and hardly feasible in case of large sample volumes.

(IV)

Amplicons of the same size might originate from different genomic loci, therefore masking polymorphisms. Hence, two bands, considered as common character of the genotypes investigated, might in fact represent polymorphic loci.

4.3 Future prospects

Provided that the species of interest contains a sufficient amount of evolutionarily young SINE copies, the ISAP is a highly convenient marker system for genotyping purposes in plant breeding, for example cultivar identification (Seibt *et al.*, 2012). Like other PCR-based markers it represents a quick, simple and cost-effective technique without the disadvantage of less reproducibility by use of unspecific primers (e.g. RAPD and ISSR) or the laborious and time-consuming primer development (e.g. SSR) (McGregor *et al.*, 2000; Nybom, 2004). Compared to contemporary high-throughput sequencing-based marker systems (Nybom *et al.*, 2014), low cost approaches like SSRs have still overweighted the relatively low number of polymorphic loci.

Retrotransposons play a key role in speciation and trigger genetic variability even among individuals within a species through lineage-specific amplification (Morgante *et al.*, 2007; Mascagni *et al.*, 2017; Serrato-Capuchina and Matute, 2018). Among retrotransposon-based markers, SINEs are especially suitable for marker development due to the easy and fast detection (Wenke *et al.*, 2011), also from short sequencing reads (Chapter 3.3, Figure 1).

In potato, many favorable circumstances for an ISAP establishment coincided: the availability of a genome reference assembly (The Potato Genome Sequencing Consortium, 2011), uncomplicated DNA extraction together with presumably still active SINE families (SolS-IIIa and SolS-IV) (Seibt *et al.*, 2016) enabled a highly efficient differentiation of potato cultivars, providing even higher resolution than SSRs (Reid *et al.*, 2009; Reid *et al.*, 2011; Seibt *et al.*, 2012).

Furthermore, Seibt *et al.* (2012) confirmed the value of retrotransposon-based markers for the detection of heritable somaclonal variations (Campbell *et al.*, 2011; Osipova *et al.*, 2011). Previous attempts to distinguish tissue culture regenerants using RAPD, SSR and AFLP often exhibit low polymorphism rates (Guimaraes *et al.*, 2009; Perrini *et al.*, 2009). As transposable elements are mainly responsible for somatic mutations (Grandbastien *et al.*, 1989; Hirochika, 1993; Huang *et al.*, 2009), their application might still be useful for polymorphism detection after *in vitro* culture, as long as highly informative high throughput sequencing techniques (Carrier *et al.*, 2012) are not affordable for breeding institutions. Tissue culture further induces alterations in the DNA methylation pattern, which

are not necessarily stable (Dann and Wilson, 2011; Baránek *et al.*, 2015). However, methylation-sensitive markers also proved their ability for the detection of somatic variations (Schellenbaum *et al.*, 2008; Baránek *et al.*, 2016) and can be combined with TE markers (Bobadilla Landey *et al.*, 2015).

A major drawback of ISAP is the requirement of genomic DNA consisting of fragments larger than ~ 5 kb, also described for other multi-locus PCR-based markers like RAPD and ISSR (Silva *et al.*, 2006; Sá *et al.*, 2011). Single-locus approaches like SSR markers might still provide results in case of partly degraded DNA samples, since the respective microsatellite arrays are usually not larger than 1 kb, rather less (Abdurakhmonov, 2016). However, SSRs are rapidly evolving loci and might not precisely mirror the underlying genomic relatedness (Guichoux *et al.*, 2011), while SINE insertions are irreversible and the ancestral state can be traced by the respective 'empty sites' (Yadav *et al.*, 2012; Keidar *et al.*, 2018), which is especially relevant for parentage analyses. Hence, to avoid the disadvantages associated with multi-locus analyses, while maintaining the benefits of the SINE-based markers, locus-specific markers might be derived from ISAP profiles.

Sequence characterized amplified regions (SCAR) were originally derived from highly variable, diagnostic bands of RAPD patterns (Paran and Michelmore, 1993) and have proven their utility, for example in cultivar identification (Turkec *et al.*, 2006) or in detection of somaclones (Osipova *et al.*, 2011). Genomic regions with increased mutation frequency, so-called 'hot spots' of DNA instability, have been proposed to explain the occurrence of highly variable bands able to distinguish between highly similar genotypes (Linacero *et al.*, 2000). The development of retrotransposon-based SCAR markers might be especially advantageous for the discrimination between recently emerged hybrids and the contributing parent genotypes (e.g. hybrid larch *Larix* × *eurolepis*, hybrid poplars), as interspecific hybridization is associated with massive mobilization of TEs (Madlung and Comai, 2004; Senerchia *et al.*, 2015). Hence, insertion polymorphisms of differentially amplified SINEs might strongly enhance ISAP resolution.

Genotypes of the *L.* × *eurolepis* parent species European larch (*Larix decidua*) and their intraspecific crossbred offspring showed less polymorphism using ISAP (Chapter 3.3, Figure 2). The combination of ISAP and AFLP method, the inter-SINE-restriction site polymorphism (ISRAP), together with the more sensitive capillary electrophoretic separation of amplicons enabled the differentiation of the two

L. decidua genotypes 'Tharandt' and 'breeding no.91'. The application of two PinS-II-derived SINE primers in combination with the *Eco*RI adapter primer revealed 21 and 24 polymorphic peaks of 32 and 30 total peak size classes, respectively (Chapter 3.3, Figure 3 - 6).

However, the comparative *RepeatExplorer* analysis of both *L.* × *eurolepis* parent species, European and Japanese larch (*Larix kaempferi*) revealed an uniformly composed repeat fraction with the major difference being the *L. decidua*-specific satellite EuLaSat3a (Tony Heitkam, personal communication). This is in line with previous findings that *L. kaempferi* SSR markers (Isoda and Watanabe, 2006) were inapplicable for *L. decidua* genotypes due to insufficient polymorphism (Wagner *et al.*, 2012). The repetitive genome portion of European and Japanese larch is highly similar (68 % and 69 %, respectively) as the contribution of the major repeat classes Ty1-*copia* and Ty3-*gypsy* (24 % and 31 % each). Hybrid larch genotypes were not included in *RepeatExplorer* analyses, so far (T. Heitkam, personal communication).

Non-LTR retrotransposons, comprising LINEs and SINEs, cover only marginal genome portions of gymnosperms. The highest LINE content in a gymnosperm so far was found in loblolly pine (*Pinus taeda*) with 2.35 % of the genome (Wegrzyn *et al.*, 2014). SINEs are even less frequent and account for only 0.001 %, like observed for other gymnosperms like Scots pine (*P. sylvestris*) and Norway spruce (*P. abies*) (Nystedt *et al.*, 2013). Irrespective of genome-wide abundance, differentially amplified retrotransposons in parent genomes and hybrids might constitute a source for SCAR marker development. Gymnosperms exhibit especially long introns (De La Torre *et al.*, 2014), for example of up to 158 kb in loblolly pine (average of 2.4 kb), consisting of more than 50 % of retrotransposons (Wegrzyn *et al.*, 2014). Accordingly, length polymorphisms between adjacent retrotransposons within intronic regions are highly relevant for PCR amplification.

How retrotransposon activity patterns are affected upon hybridization is not investigated in gymnosperms so far. In *Arabidopsis*, hybrid-specific alterations in TE expression, although found to be rare in F1 hybrids of *A. thaliana* and *A. lyrata*, predominantly occur near genes (Göbel *et al.*, 2018). The genome of hexaploid bread wheat (*Triticum aestivum*) emerged by two independent intraspecific hybridizations: *Triticum urartu* and *Aegilops speltoides* contributed to a tetraploid species from which the domesticated *Triticum turgidum* arose. *T. turgidum* hybridized with *Aegilops tauschii* resulting in

hexaploid *T. aestivum* (Matsuoka, 2011). Hence, the exceptional high abundance of the Au SINE in the wheat genome (Chapter 2.2, Figure 1) might be explained by those hybridization events, as two of three parent species (*T. urartu* and *Ae. tauschii*) contain only 133 and 180 Au copies, respectively (Keidar *et al.*, 2018). A significant increase in Au copy number was only detected in one of three newly formed allopolyploid wheat species (*T. turgidum* ssp. *durum* × *Ae. tauschii*; Ben-David *et al.*, 2013), indicating that SINE (or TE) bursts do not necessarily follow an interspecific hybridization (Wicker *et al.*, 2018).

Further studies on recently developed hybrids used DArT markers for the quantification of the parental influence on intergeneric hybrids of the grasses *Festuca* and *Lolium* (Kopecký *et al.*, 2011), or a combination of nuclear and chloroplast SSR markers to comparatively characterize *Magnolia* hybrids with regard to the contributing species including the assignment of morphological traits (Muranishi *et al.*, 2013).

The present tendency is towards SNP-based genome-wide association studies (GWAS) that facilitate a more targeted marker development for parentage analyses and presumably will provide a deeper insight into mechanisms underlying the merging of different species in a hybrid (Zheng *et al.*, 2017; Kim *et al.*, 2018).

References

Abdurakhmonov, I.Y.A.E.-I.Y. (2016) Introduction to microsatellites: basics, trends and highlights. In *Microsatellite markers*. Rijeka: IntechOpen.

Baránek, M., Čechová, J., Kovacs, T., Eichmeier, A., Wang, S., Raddová, J., Nečas, T. and Ye, X. (2016) Use of combined MSAP and NGS techniques to identify differentially methylated regions in somaclones: a case study of two stable somatic wheat mutants. *PLoS One*, **11**, e0165749.

Baránek, M., Čechová, J., Raddová, J., Holleinová, V., Ondrušíková, E. and Pidra, M. (2015) Dynamics and reversibility of the DNA methylation landscape of grapevine plants (*Vitis vinifera*) stressed by *in vitro* cultivation and thermotherapy. *PLoS One*, **10**, e0126638.

Ben-David, S., Yaakov, B. and Kashkush, K. (2013) Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat. *Plant J.*, **76**, 201–210.

Bobadilla Landey, R., Cenci, A., Guyot, R., et al. (2015) Assessment of genetic and epigenetic changes during cell culture ageing and relations with somaclonal variation in *Coffea arabica*. *Plant Cell, Tissue Organ Cult.*, **122**, 517–531.

Bruegmann, T. and Fladung, M. (2013) Potentials and limitations of the cross-species transfer of nuclear microsatellite marker in six species belonging to three sections of the genus *Populus* L. *Tree Genet. Genomes*, **9**, 1413–1421.

Campbell, B.C., LeMare, S., Piperidis, G. and Godwin, I.D. (2011) IRAP, a retrotransposon-based marker system for the detection of somaclonal variation in barley. *Mol. Breed.*, **27**, 193–206.

Carrier, G., Cunff, L. Le, Dereeper, A., Legrand, D., Sabot, F., Bouchez, O., Audeguin, L., Boursiquot, J.-M. and This, P. (2012) Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. *PLoS One*, **7**, e32973.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189.

Dann, A.L. and Wilson, C.R. (2011) Comparative assessment of genetic and epigenetic variation among regenerants of potato (*Solanum tuberosum*) derived from long-term nodal tissue-culture and cell selection. *Plant Cell Rep.*, **30**, 631–639.

Deragon, J.-M. and Zhang, X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, **55**, 949–956.

Farell, E.M. and Alexandre, G. (2012) Bovine serum albumin further enhances the effects of organic solvents on increased yield of polymerase chain reaction of GC-rich templates. *BMC Res. Notes*, **5**, 257.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.

Garrido-Cardenas, J.A., Mesa-Valle, C. and Manzano-Agugliaro, F. (2018) Trends in plant research using molecular markers. *Planta*, **247**, 543–557.

Göbel, U., Arce, A.L., He, F., Rico, A., Schmitz, G. and Meaux, J. de (2018) Robustness of transposable element regulation but no genomic shock observed in interspecific *Arabidopsis hybrids*. *Genome Biol. Evol.*, **10**, 1403–1415.

González-Pérez, S., Mallor, C., Garcés-Claver, A., Merino, F., Taboada, A., Rivera, A., Pomar, F., Perovic, D. and Silvar, C. (2015) Exploring genetic diversity and quality traits in a collection of onion (*Allium cepa* L.) landraces from north-west Spain. *Genetika*, **47**, 885–900.

Grandbastien, M.-A., Spielmann, A. and Caboche, M. (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*, **337**, 376–380.

Guichoux, E., Lagache, L., Wagner, S., et al. (2011) Current trends in microsatellite genotyping. *Mol. Ecol. Resour.*, **11**, 591–611.

Guimaraes, N., Torga, P., Resende, E. De, Chalfun, A., Paiva, E. and Paiva, L. (2009) Identification of somaclonal variants in 'Prata Ana' banana using molecular and cytogenetic techniques [In Portuguese]. *Cienc. e Agrotecnologia*, **33**, 448–454.

Henry, R.J. (2001) Plant DNA extraction. In R. J. Henry, ed. *Plant Genotyping: the DNA fingerprinting of plants.* United Kingdom: CAB International, pp. 239–249.

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

Huang, H., Tong, Y., Zhang, Q. and Gao, L. (2013) Genome size variation among and within *Camellia* species by using flow cytometric analysis. *PLoS One*, **8**, e64981.

Huang, J., Zhang, K., Shen, Y., Huang, Z., Li, M., Tang, D., Gu, M. and Cheng, Z. (2009) Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice. *Genomics*, **93**, 274–281.

Jiang, G.-L. (2013) Molecular markers and marker-assisted breeding in plants. In S. B. Andersen, ed.

Plant breeding from laboratories to fields. Rijeka: IntechOpen.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V and Jurka, M. V (2005) Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. *Cytogenet. Genome Res.*, **110**, 117–123.

Katterman, F.R.H. and Shattuck, V.I. (1983) An effective method of DNA isolation from the mature leaves of *Gossypium* species that contain large amounts of phenolic terpenoids and tannins. *Prep Biochem*, **14**, 347–359.

Keidar, D., Doron, C. and Kashkush, K. (2018) Genome-wide analysis of a recently active retrotransposon, Au SINE, in wheat: content, distribution within subgenomes and chromosomes, and gene associations. *Plant Cell Rep.*, **37**, 193–208.

Kim, B., Udvardi, M.K., Zhang, W., et al. (2018) GWASpro: a high-performance genome-wide association analysis server. *Bioinformatics*, bty989, doi: 10.1093/bioinformatics/bty989

Kopecký, D., Bartoš, J., Christelová, P., Černoch, V., Kilian, A. and Doležel, J. (2011) Genomic constitution of *Festuca* × *Lolium* hybrids revealed by the DArTFest array. *Theor. Appl. Genet.*, **122**, 355–363.

Kramer, M.F. and Coen, D.M. (2006) Enzymatic amplification of DNA by PCR: standard procedures and optimization. *Curr. Protoc. Cytom.*, **37**, A.3K.1–A.3K.15.

La Torre, A.R. De, Birol, I., Bousquet, J., et al. (2014) Insights into conifer giga-genomes. *Plant Physiol.*, **166**, 1724–1732.

Lenoir, A., Lavie, L., Prieto, J.L., Goubely, C., Coté, J.C., Pélissier, T. and Deragon, J.M. (2001) The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **18**, 2315–2322.

Lin, Y.-C., Wang, J., Delhomme, N., et al. (2018) Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proc. Natl. Acad. Sci.*, **115**, E10970–E10978.

Linacero, R., Freitas Alves, E. and Vázquez, A.M. (2000) Hot spots of DNA instability revealed through the study of somaclonal variation in rye. *Theor. Appl. Genet.*, **100**, 506–511.

Liu, X., Wang, Z., Wang, D. and Zhang, J. (2016) Phylogeny of *Populus-Salix* (Salicaceae) and their relative genera using molecular datasets. *Biochem. Syst. Ecol.*, **68**, 210–215.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) Reverse transcription of

R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.

Madlung, A. and Comai, L. (2004) The effect of stress on genome regulation and structure. *Ann. Bot.*, **94**, 481–495.

Mascagni, F., Giordani, T., Ceccarelli, M., Cavallini, A. and Natali, L. (2017) Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). *BMC Genomics*, **18**, 634.

Matsuoka, Y. (2011) Evolution of polyploid *Triticum* wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol.*, **52**, 750–764.

McGregor, C.E., Lambert, C.A., Greyling, M.M., Louw, J.H. and Warnich, L. (2000) A comparative assessment of DNA fingerprinting techniques (RAPD, ISSR, AFLP and SSR) in tetraploid potato (*Solanum tuberosum* L.) germplasm. *Euphytica*, **113**, 135–144.

Moreira, P. and Oliveira, D. (2011) Leaf age affects the quality of DNA extracted from *Dimorphandra mollis* (Fabaceae), a tropical tree species from the Cerrado region of Brazil. *Genet Mol Res*, **10**, 353–358.

Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with non-repetitive DNA in plant genomes. *Nat. Genet.*, **30**, 194.

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

Muranishi, S., Tamaki, I., Setsuko, S. and Tomaru, N. (2013) Asymmetric introgression between *Magnolia stellata* and *M. salicifolia* at a site where the two species grow sympatrically. *Tree Genet. Genomes*, **9**, 1005–1015.

Myouga, F., Tsuchimoto, S., Noma, K., Ohtsubo, H. and Ohtsubo, E. (2001) Identification and structural analysis of SINE elements in the *Arabidopsis thaliana* genome. *Genes Genet. Syst.*, **76**, 169–179.

Nybom, H. (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.*, **13**, 1143–1155.

Nybom, H., Weising, K. and Rotter, B. (2014) DNA fingerprinting in botany: past, present, future. *Investig. Genet.*, **5**, 1.

Nystedt, B., Street, N.R., Wetterbom, A., et al. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579.

Okada, N. (1991) SINEs. *Curr. Opin. Genet. Dev.*, **1**, 498–504.

Osipova, E.S., Lysenko, E.A., Troitsky, A. V, Dolgikh, Y.I., Shamina, Z.B. and Gostimskii, S.A. (2011) Analysis of SCAR marker nucleotide sequences in maize (*Zea mays* L.) somaclones. *Plant Sci.*, **180**, 313–322.

Ostrowska, E., Muralitharan, M., Chandler, S., Volker, P., Hetherington, S. and Dunshea, F. (1998) Optimizing conditions for DNA isolation from *Pinus radiata*. *Vitr. Cell. Dev. Biol. - Plant*, **34**, 108–111.

Paran, I. and Michelmore, R.W. (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *Theor. Appl. Genet.*, **85**, 985–993.

Perrini, R., Alba, V., Ruta, C., Morone-Fortunato, I., Blanco, A. and Montemurro, C. (2009) An evaluation of a new approach to the regeneration of *Helichrysum italicum* (Roth) G. Don, and the molecular characterization of the variation among sets of differently derived regenerants. *Cell. Mol. Biol. Lett.*, **14**, 377–394.

Reid, A., Hof, L., Esselink, D. and Vosman, B. (2009) Potato cultivar genome analysis. In R. Burns, ed. *Plant pathology: techniques and protocols.* Totowa, NJ: Humana Press, pp. 295–308.

Reid, A., Hof, L., Felix, G., et al. (2011) Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue. *Euphytica*, **182**, 239.

Reid, A. and Kerr, E.M. (2007) A rapid simple sequence repeat (SSR)-based identification method for potato cultivars. *Plant Genet. Resour.*, **5**, 7–13.

Sá, O., Pereira, J.A. and Baptista, P. (2011) Optimization of DNA extraction for RAPD and ISSR analysis of *Arbutus unedo* L. leaves. *Int. J. Mol. Sci.*, **12**, 4156–4164.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.*, **81**, 8014–8018.

Schellenbaum, P., Mohler, V., Wenzel, G. and Walter, B. (2008) Variation in DNA methylation patterns of grapevine somaclones (*Vitis vinifera* L.). *BMC Plant Biol.*, **8**, 78.

Schwichtenberg, K., Wenke, T., Zakrzewski, F., Seibt, K.M., Minoche, A., Dohm, J.C.,

Weisshaar, B., Himmelbauer, H. and Schmidt, T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229–244.

Seibt, K.M., Wenke, T., Muders, K., Truberg, B. and Schmidt, T. (2016) Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. *Plant J.*, **86**, 268–285.

Seibt, K.M., Wenke, T., Wollrab, C., Junghans, H., Muders, K., Dehmer, K.J., Diekmann, K. and Schmidt, T. (2012) Development and application of SINE-based markers for genotyping of potato varieties. *Theor. Appl. Genet.*, **125**, 185–196.

Senerchia, N., Felber, F. and Parisod, C. (2015) Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation. *Proc. R. Soc. B Biol. Sci.*, **282**, 20142874.

Serrato-Capuchina, A. and Matute, D.R. (2018) The role of transposable elements in speciation. *Genes*, **9**, 254.

Shepherd, M., Cross, M., Stokoe, R.L., Scott, L.J. and Jones, M.E. (2002) High-throughput DNA extraction from forest trees. *Plant Mol. Biol. Report.*, **20**, 425.

Silva, M., Ming, L., Pereira, A., Bertoni, B., Batistini, A. and Pereira, P. (2006) Phytochemical and genetic variability of *Casearia sylvestris* Sw. from São Paulo State Atlantic Forest and Cerrado populations. *Rev. Bras. Plantas Med.*, **8**, 159–166.

Tibbits, J.F.G., McManus, L.J., Spokevicius, A. V and Bossinger, G. (2006) A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. *Plant Mol. Biol. Report.*, **24**, 81–91.

Turkec, A., Sayar, M. and Heinze, B. (2006) Identification of sweet cherry cultivars (*Prunus avium* L.) and analysis of their genetic relationships by chloroplast sequence-characterised amplified regions (cpSCAR). *Genet. Resour. Crop Evol.*, **53**, 1635–1641.

Varshney, R.K., Graner, A. and Sorrells, M.E. (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.*, **23**, 48–55.

Verbylaite, R., Beišys, P., Rimas, V. and Kuusiene, S. (2010) Comparison of ten DNA extraction protocols from wood of European aspen (*Populus tremula* L.). *Balt. For.*, **16**, 35–42.

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

Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., et al. (2014) Unique features of the loblolly pine *Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, **196**, 891–909.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.

Wenke, T., Seibt, K.M., Döbel, T., Muders, K. and Schmidt, T. (2015) Inter-SINE Amplified Polymorphism (ISAP) for rapid and robust plant genotyping. In J. Batley, ed. *Plant genotyping: methods and protocols*. New York: Springer, pp. 183–192.

Wicker, T., Gundlach, H., Spannagl, M., et al. (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.*, **19**, 103.

Woide, D., Zink, A. and Thalhammer, S. (2010) Technical note: PCR analysis of minimum target amount of ancient DNA. *Am. J. Phys. Anthropol.*, **142**, 321–327.

Yadav, V.P., Mandal, P.K., Bhattacharya, A. and Bhattacharya, S. (2012) Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*. *Nat. Commun.*, **3**, 854.

Yin, T.M., Zhang, X.Y., Gunter, L.E., Li, S.X., Wullschleger, S.D., Huang, M.R. and Tuskan, G.A. (2009) Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. *New Phytol.*, **181**, 498–503.

Yoon, I., Park, D., Kim, J., et al. (2017) Identification of the biologically active constituents of *Camellia japonica* leaf and anti-hyperuricemic effect *in vitro* and *in vivo*. *Int J Mol Med*, **39**, 1613–1620.

Zheng, M., Peng, C., Liu, H., et al. (2017) Genome-wide association study reveals candidate genes for control of plant height, branch initiation height and branch number in rapeseed (*Brassica napus* L.). *Front. Plant Sci.*, 8, 1246.

Zonneveld, B.J.M. (2012) Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nord. J. Bot.*, **30**, 490–502.

Supplementary Chapter

Supplemental Information to

Chapter 2.1

The identification of SINEs in *Camellia japonica* and the multistage concept for SINE family and subfamily classification

Content

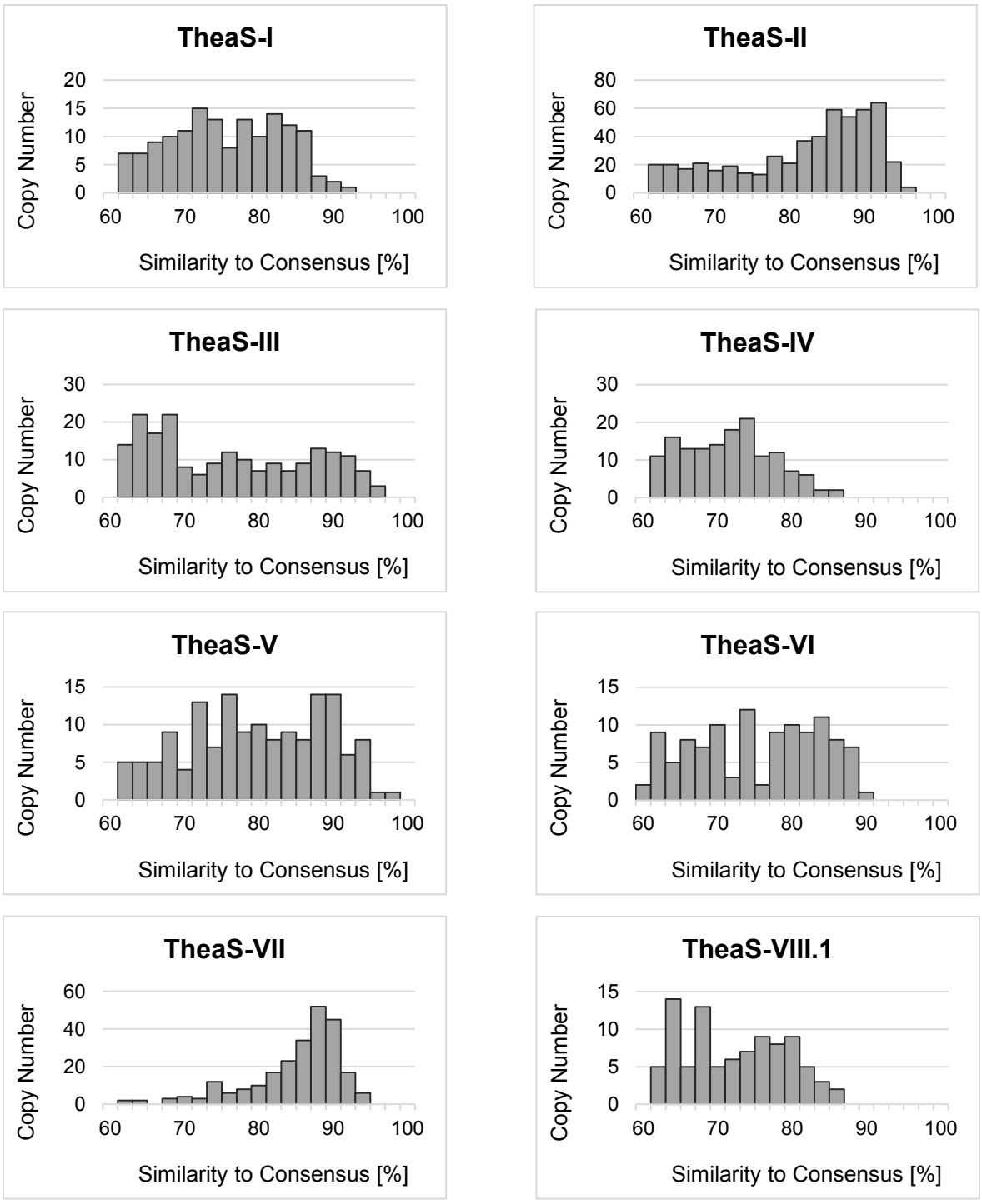


Figure S1. Activity profiles of TheaS families and subfamilies. The SINE full-length copies were pairwise compared with the consensus sequence of the respective family and subfamily. The resulting percentage identity values were assigned to similarity intervals.

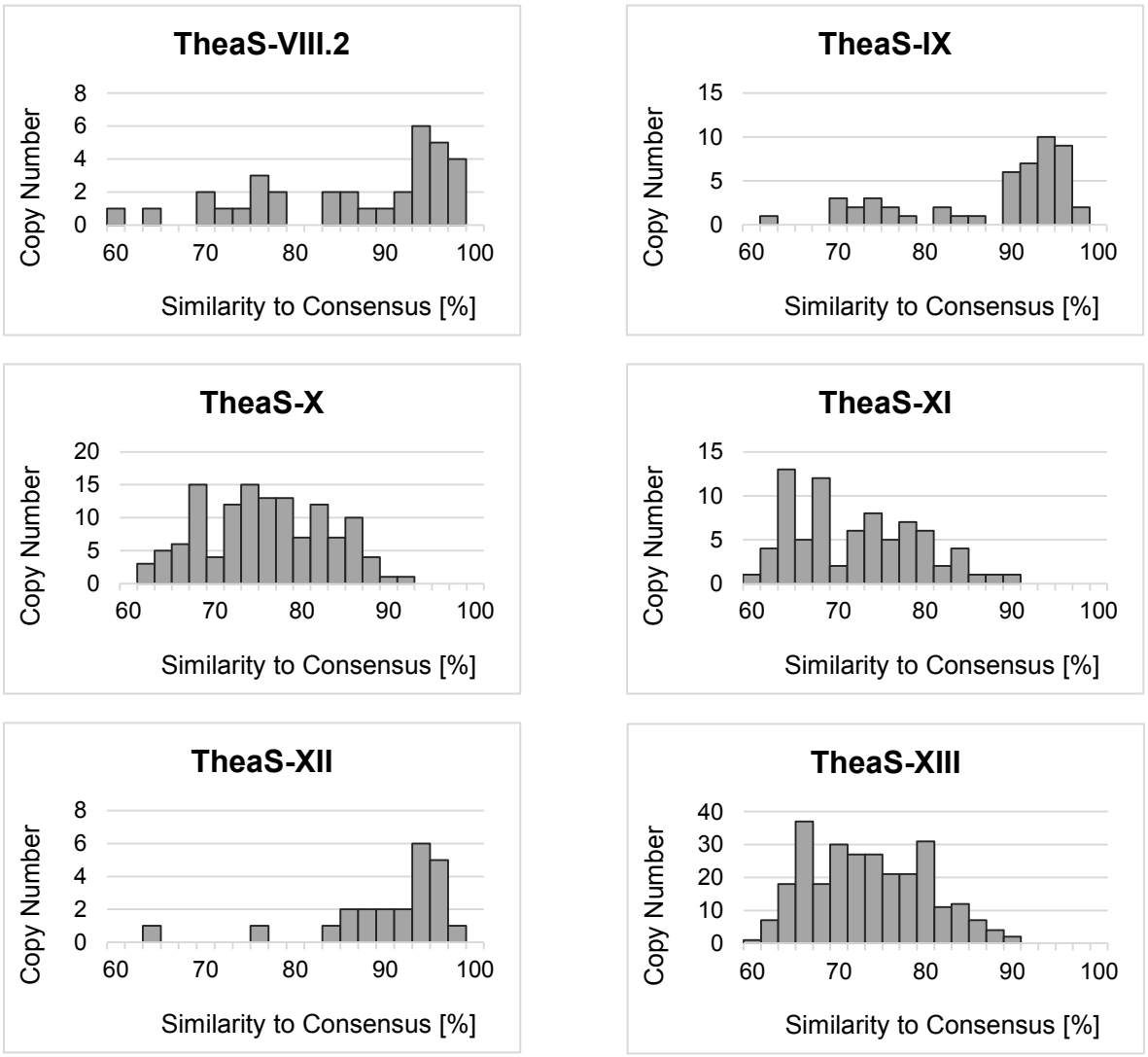


Figure S1. Continued.

Supplemental Information to

Chapter 2.2

Evolutionary modes of SINE family emergence in grasses

Content

Supplemental Figures

Supplemental Tables

Supplemental Figures

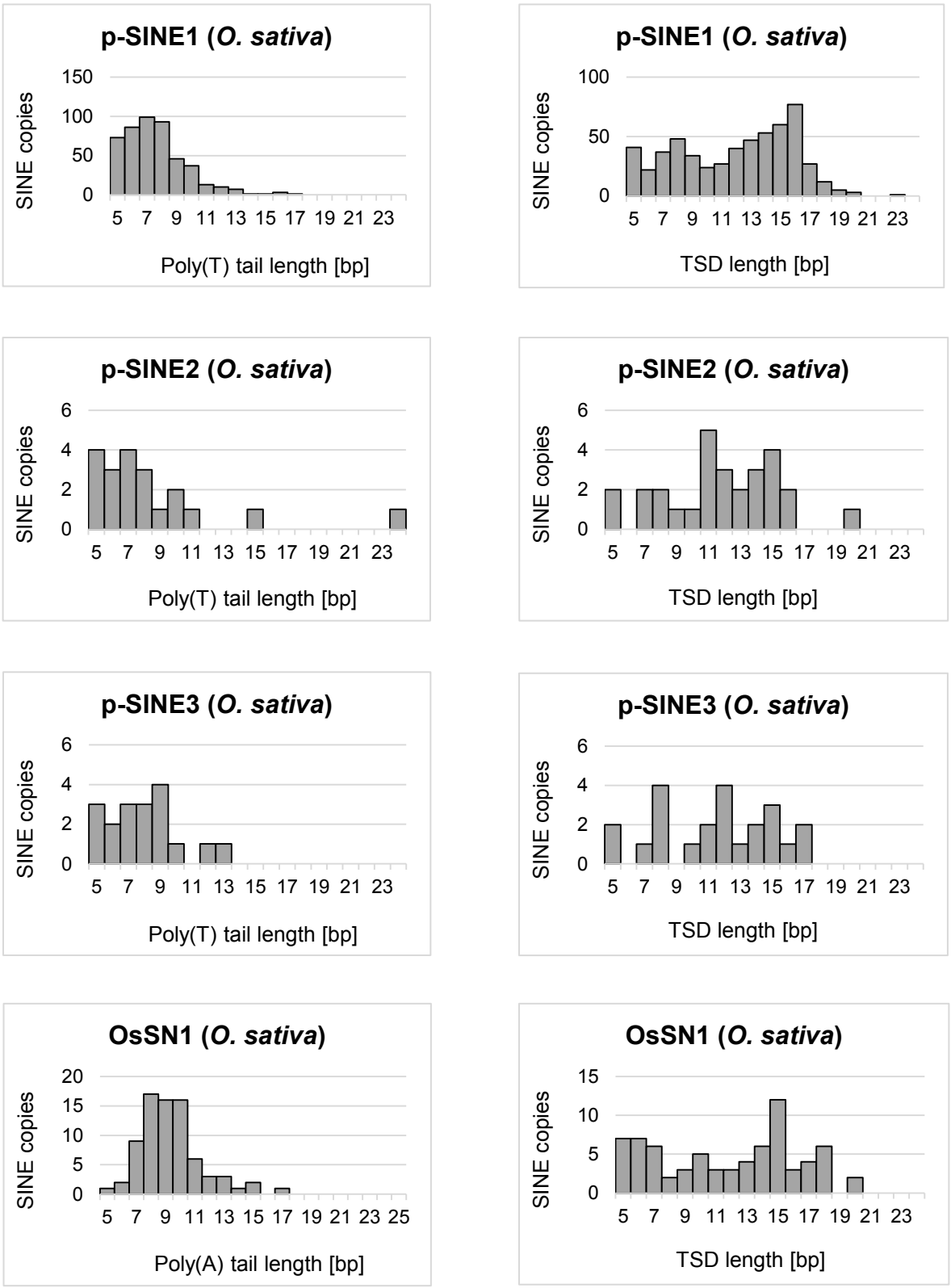
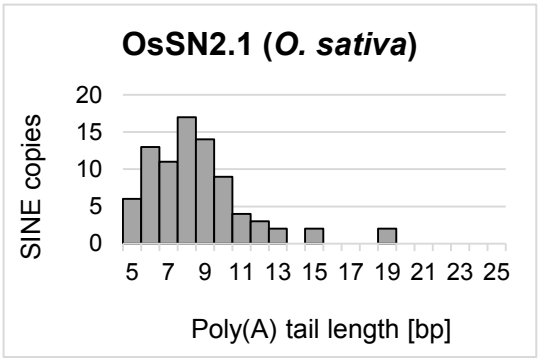
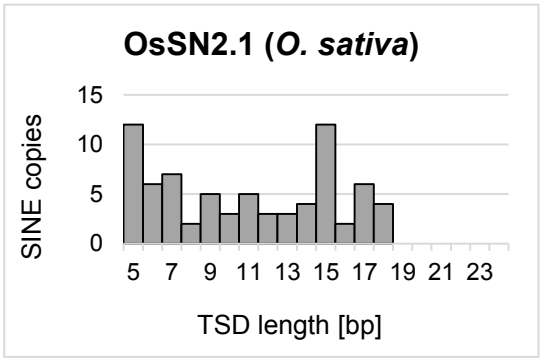
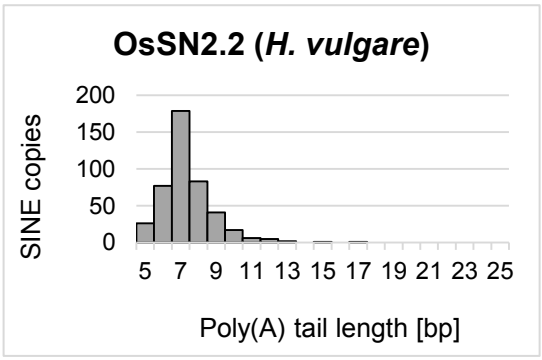
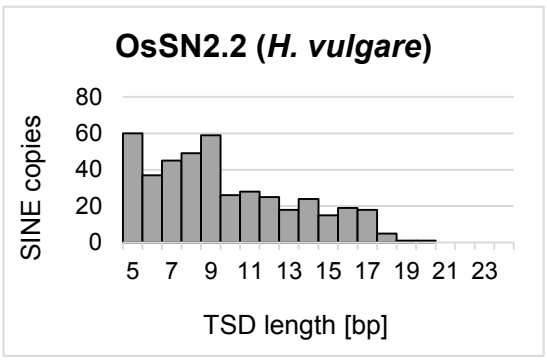


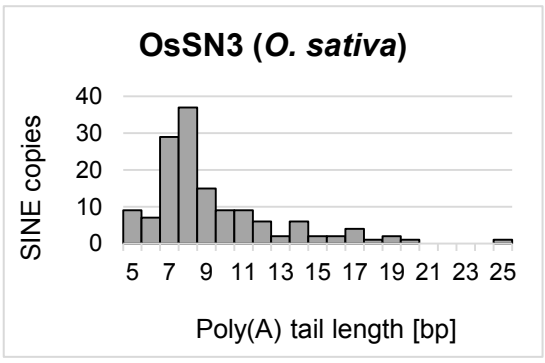
Figure S1. Lengths of 3' tails (left) and target site duplications (right) of Poaceae full-length SINE copies.

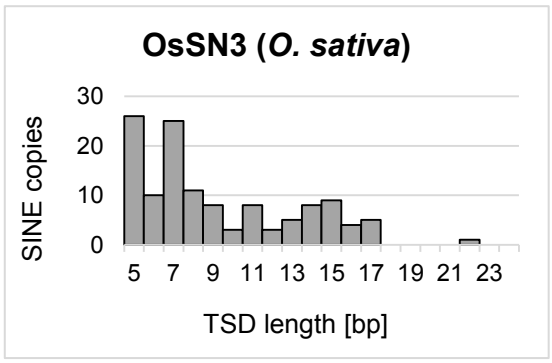


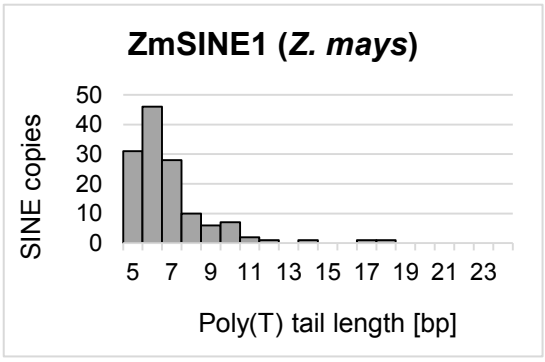












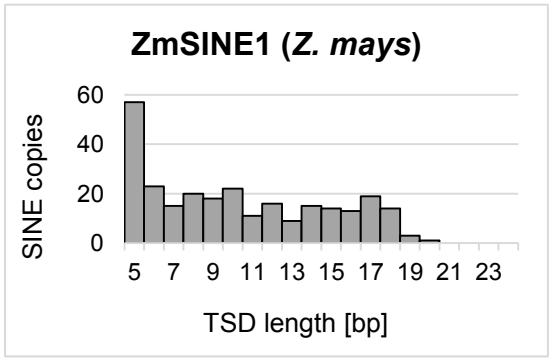
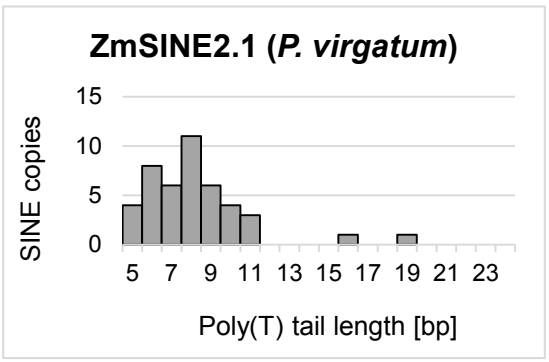
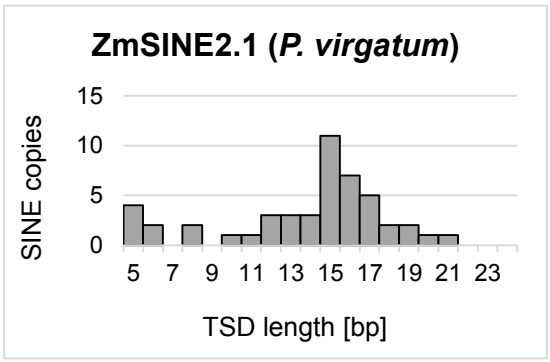
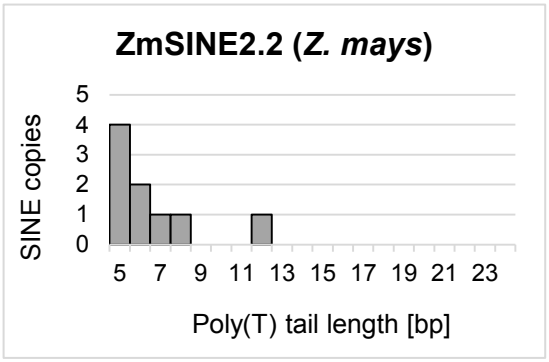
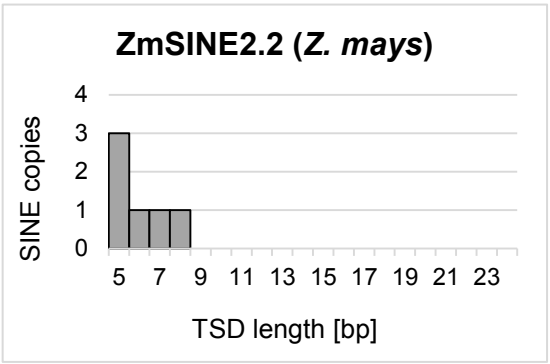


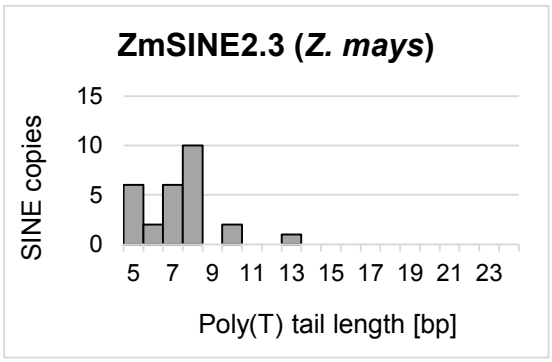
Figure S1. Continued.

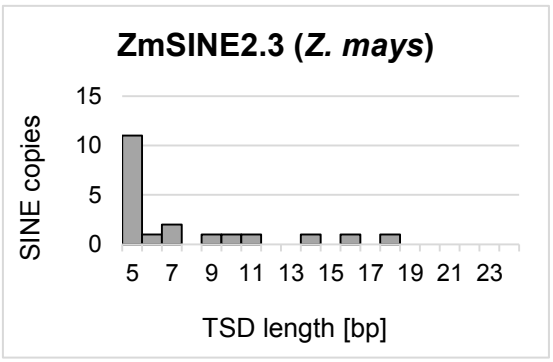


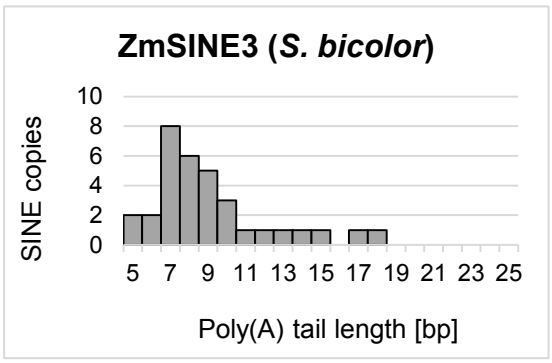












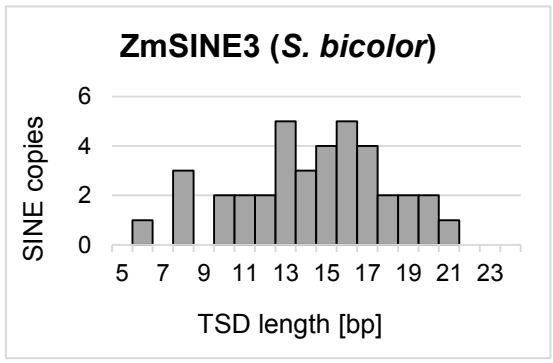
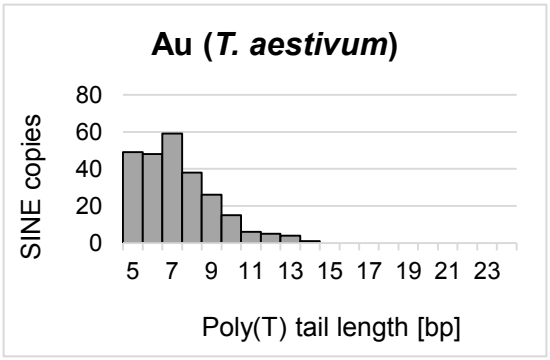
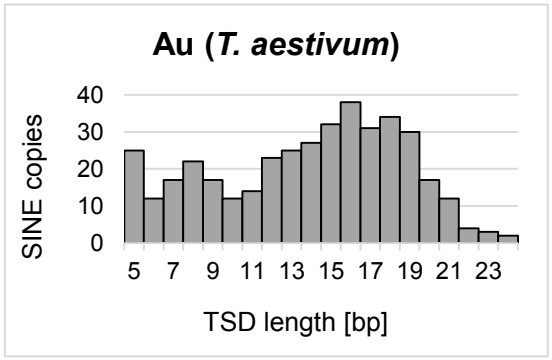
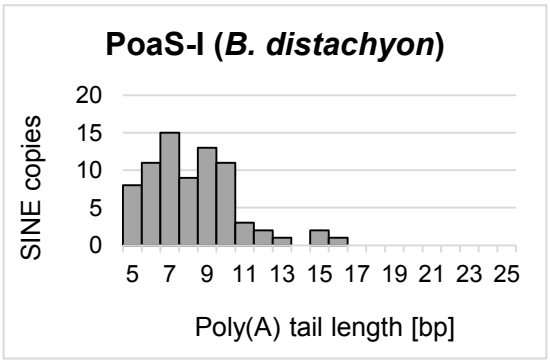
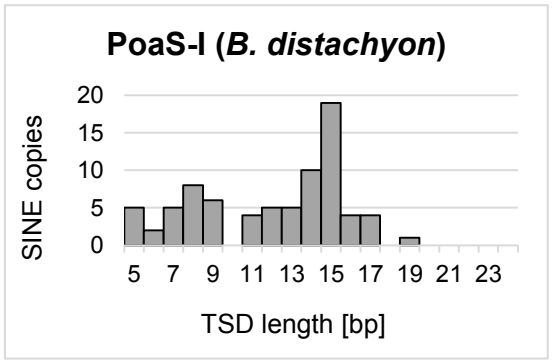


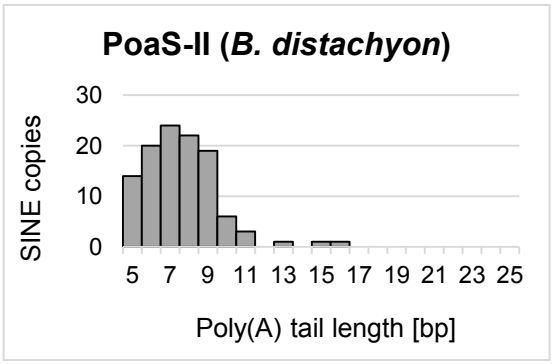
Figure S1. Continued.

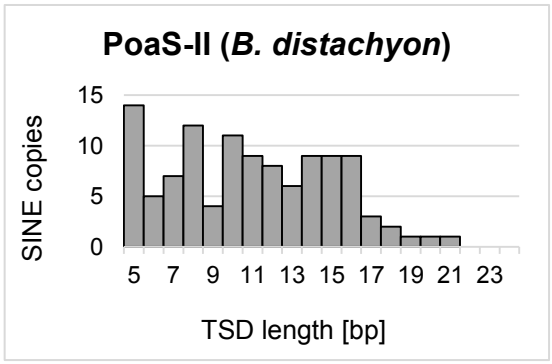


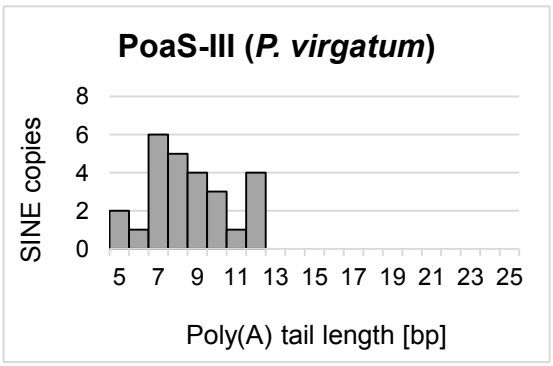












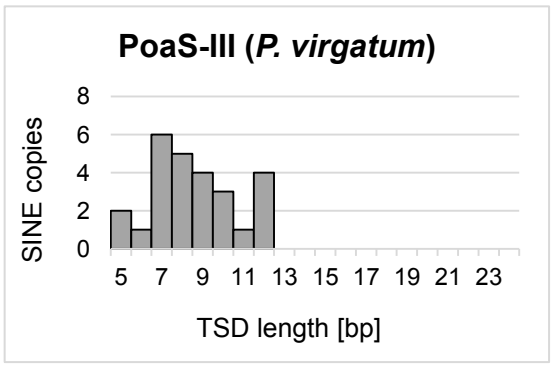
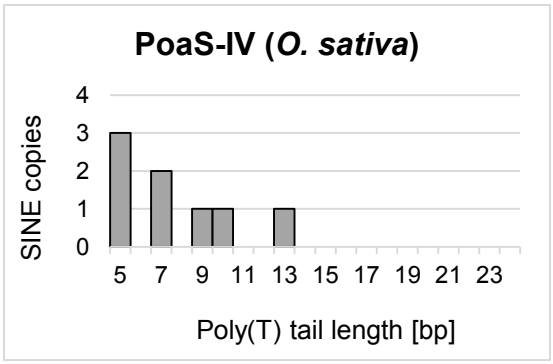
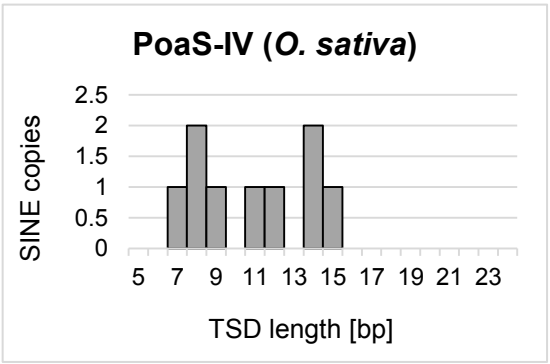
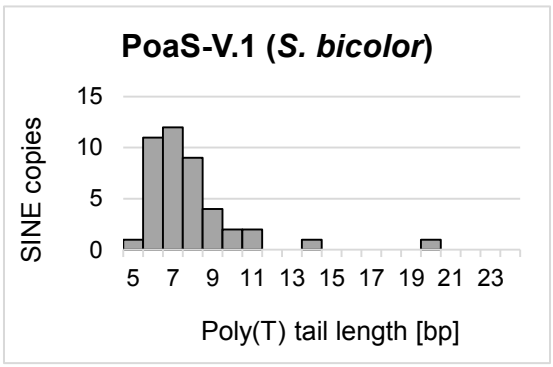
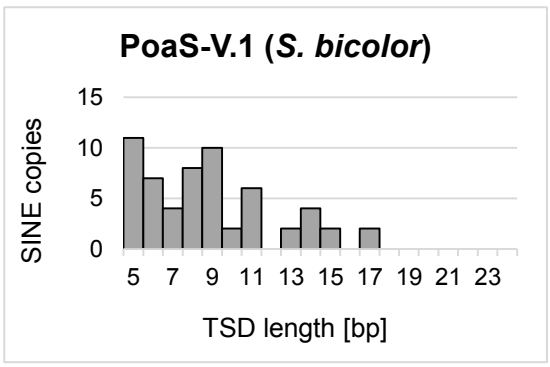


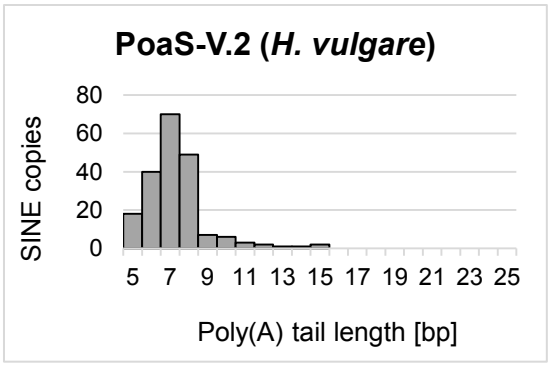
Figure S1. Continued.

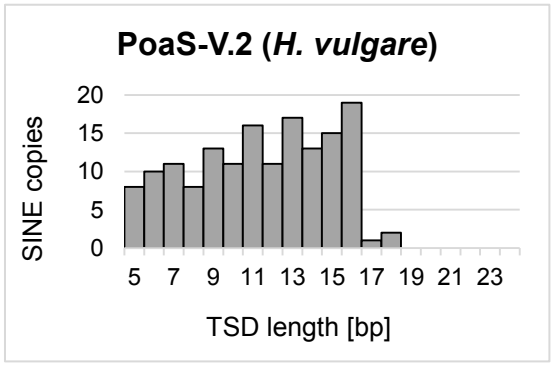


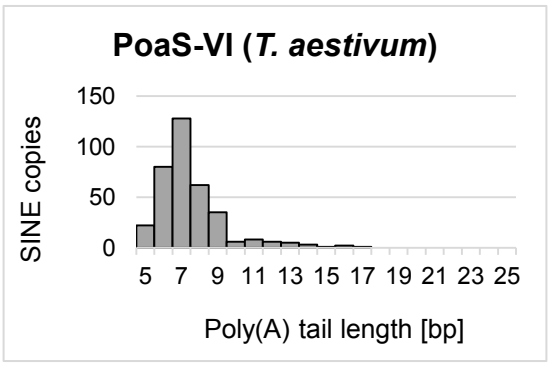












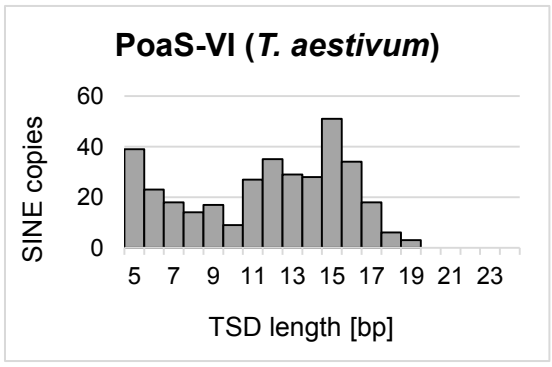
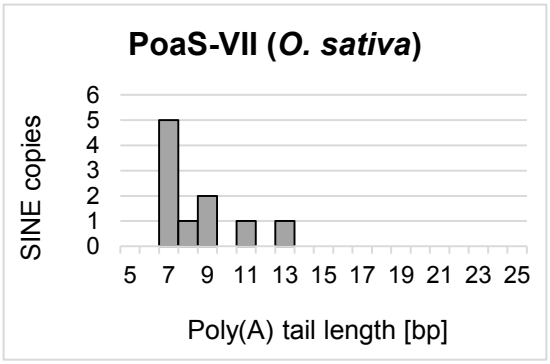
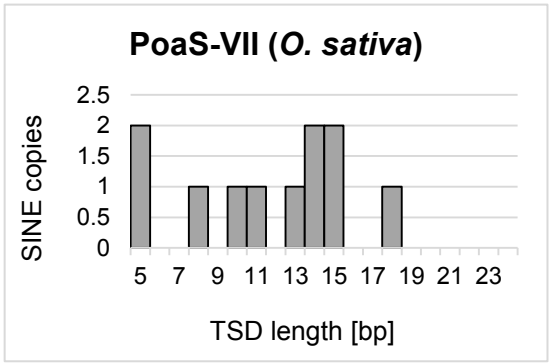
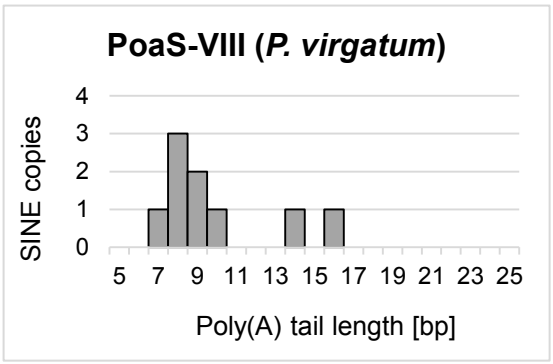
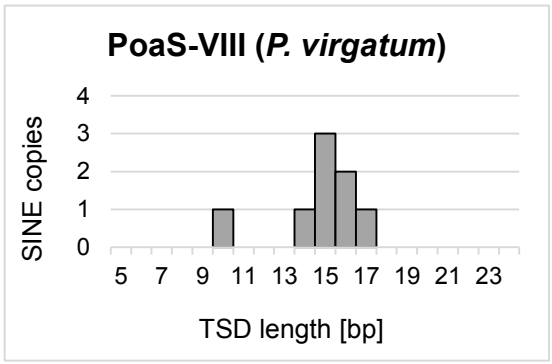


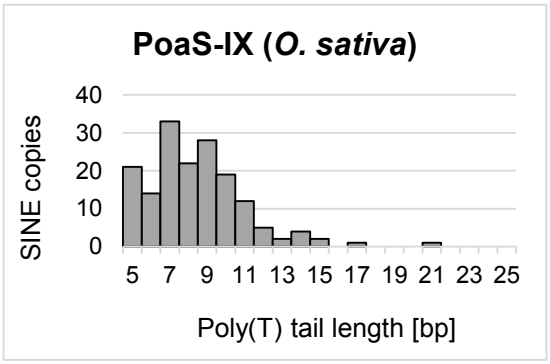
Figure S1. Continued.

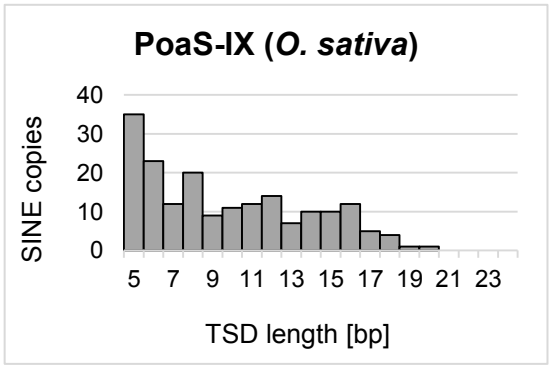


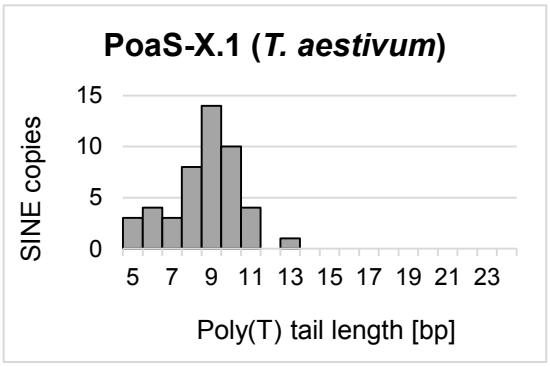












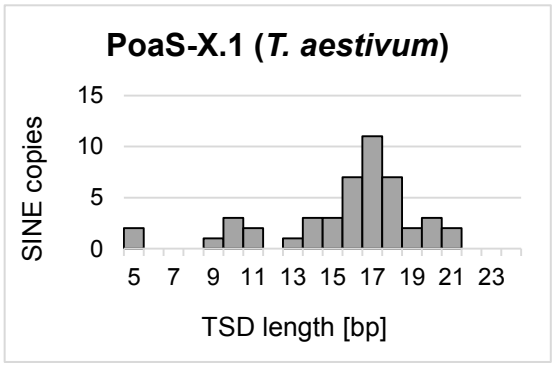
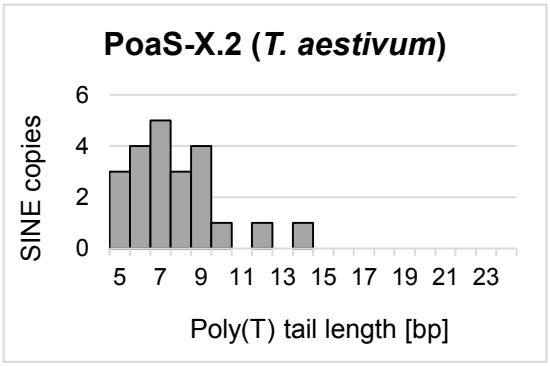
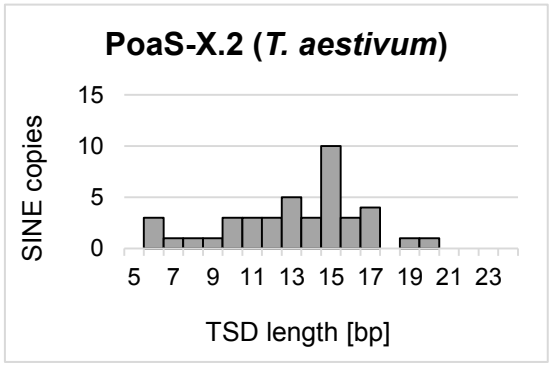
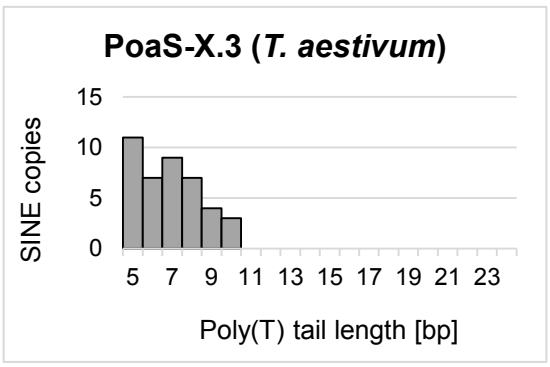
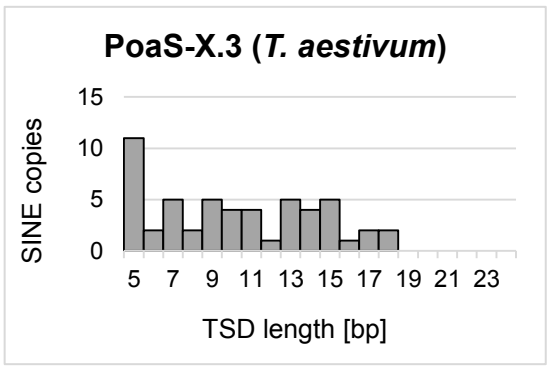


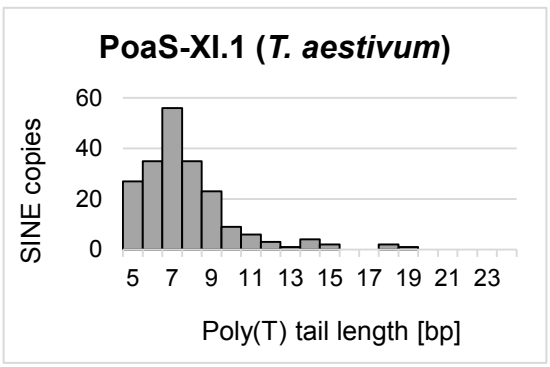
Figure S1. Continued.

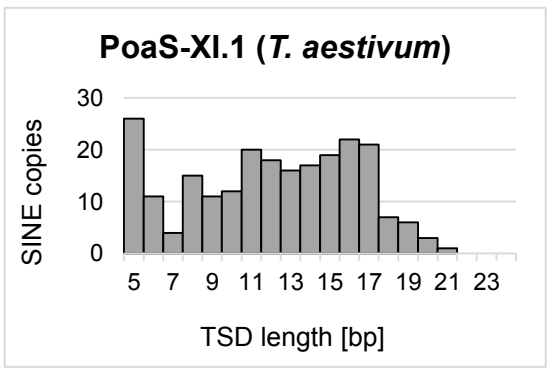


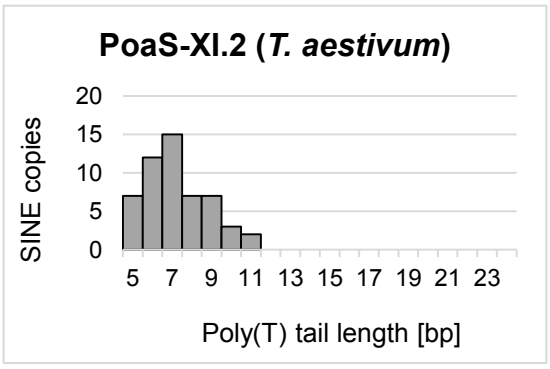












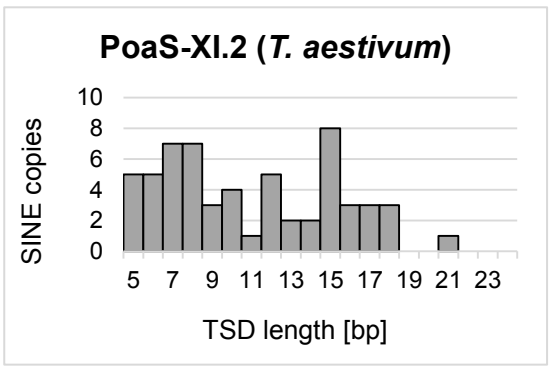


Figure S1. Continued.

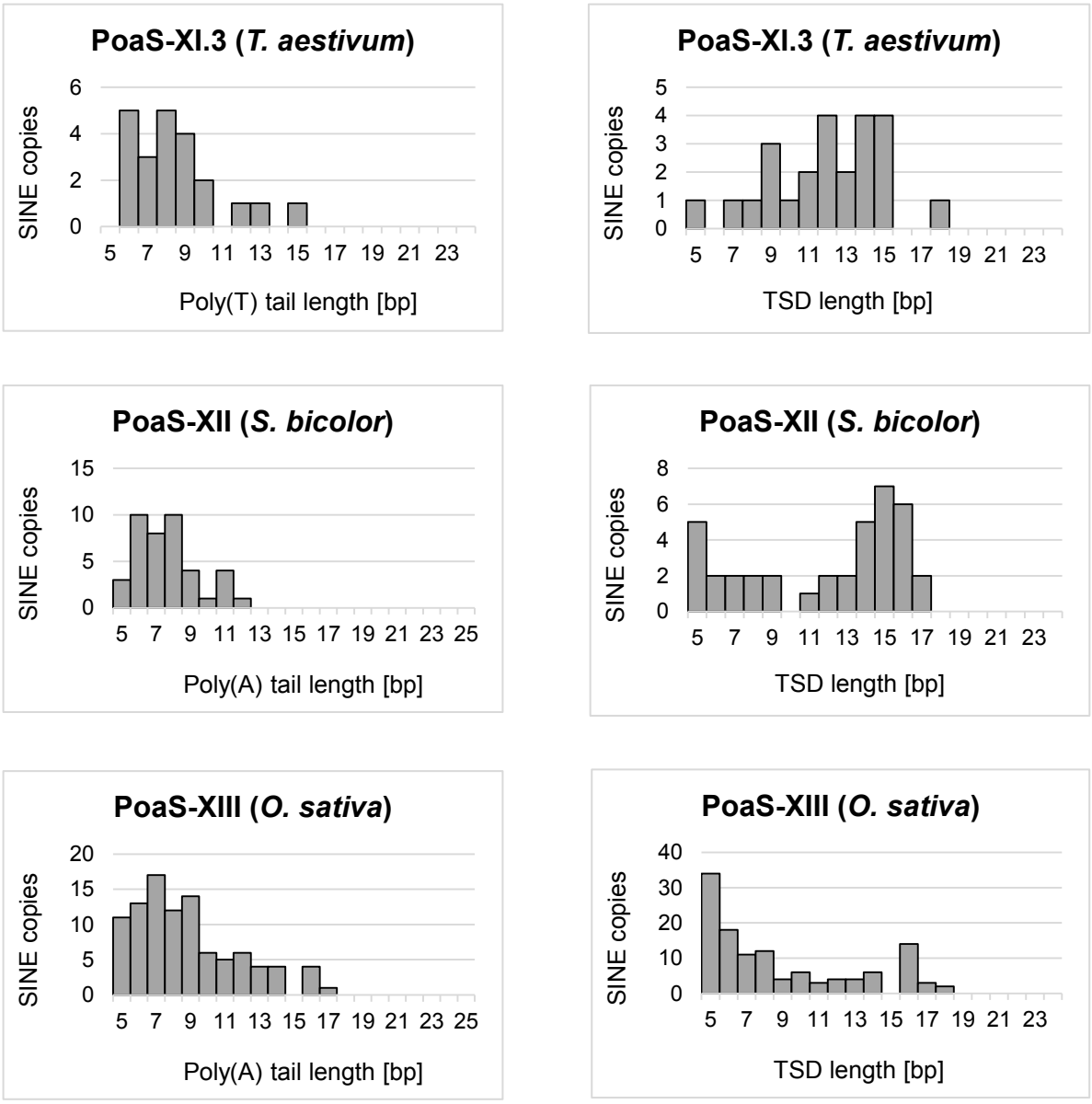


Figure S1. Continued.

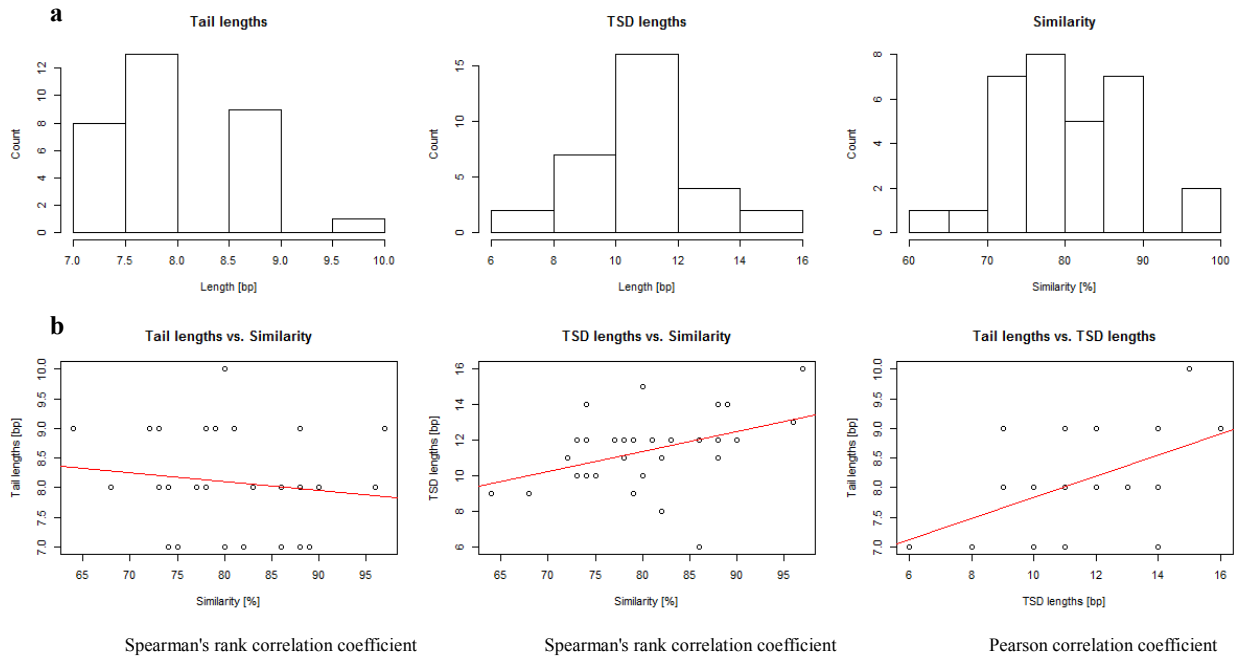


Figure S2. Correlation between target site duplication (TSD) length, tail length and similarity. (a) Histograms showing the data distribution of tail lengths, TSD lengths and similarity values. The TSD lengths and the similarity values were normally distributed ($p = 0.23$ and $p = 0.87$, respectively), while the tail lengths values were not normally distributed ($p = 0.001$). (b) Scatter plots illustrate a potential correlation between the three SINE characteristics. The red line represents the regression line. A correlation with a positive correlation factor (0.42) was detected between the TSD lengths and similarity values ($p = 0.01$). Accordingly, a positive correlation (factor = 0.40) between tail lengths and TSD lengths ($p = 0.02$) was calculated. In contrast, a negative correlation (factor = -0.18) was detected between similarity values and tail lengths ($p = 0.32$). The high p-value indicates that no significant correlation exists between similarity values and tail lengths.

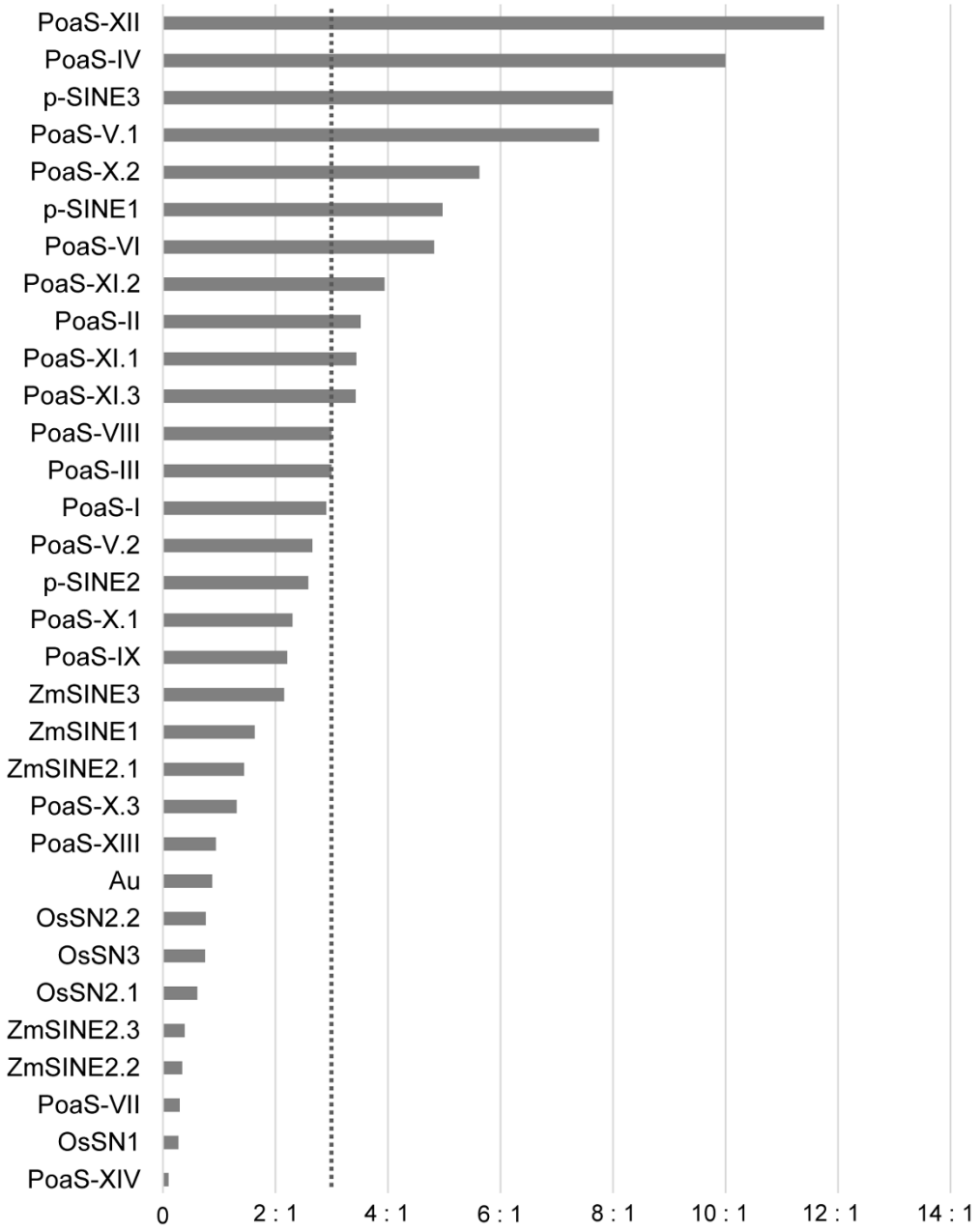


Figure S3. Ratio of full-length to 5' truncated copies. The average and median ratio of all 32 Poaceae SINE families and subfamilies (3:1) is indicated as a dotted line.

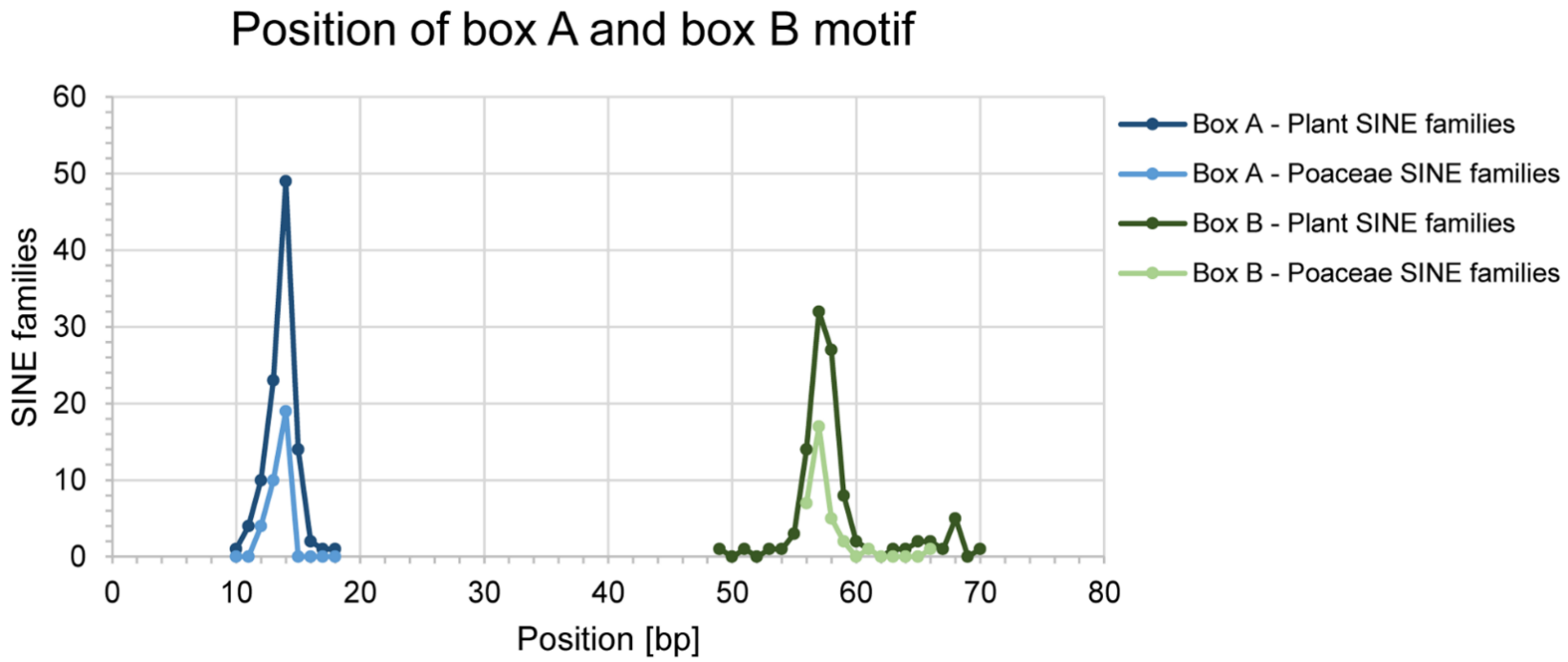


Figure S4. Position of box A and box B motif and their distance within plant SINE consensus sequences. The position of the first nucleotide of box A and box B, respectively, was determined for all 32 Poaceae SINE families and subfamilies and for 103 plant SINE families (Poaceae SINE families and subfamilies included) (Table S7).

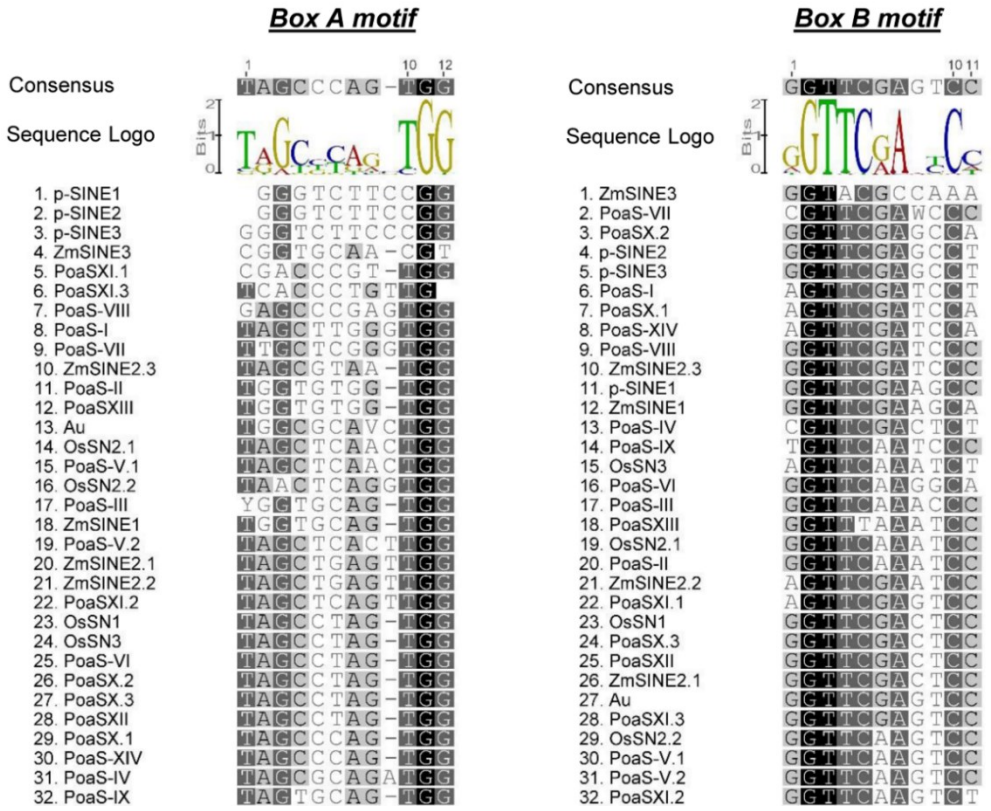


Figure S5. Conserved nucleotides of promotor motifs for Poaceae SINE families and subfamilies. The box A and B motifs of 32 Poaceae SINE families and subfamilies are shown with the consensus sequence and the respective sequence logo. Similarity shadings: black – 100 %; dark grey – 80 % to 100 %; pale grey – 60 % to 80 %, white – less than 60 %.

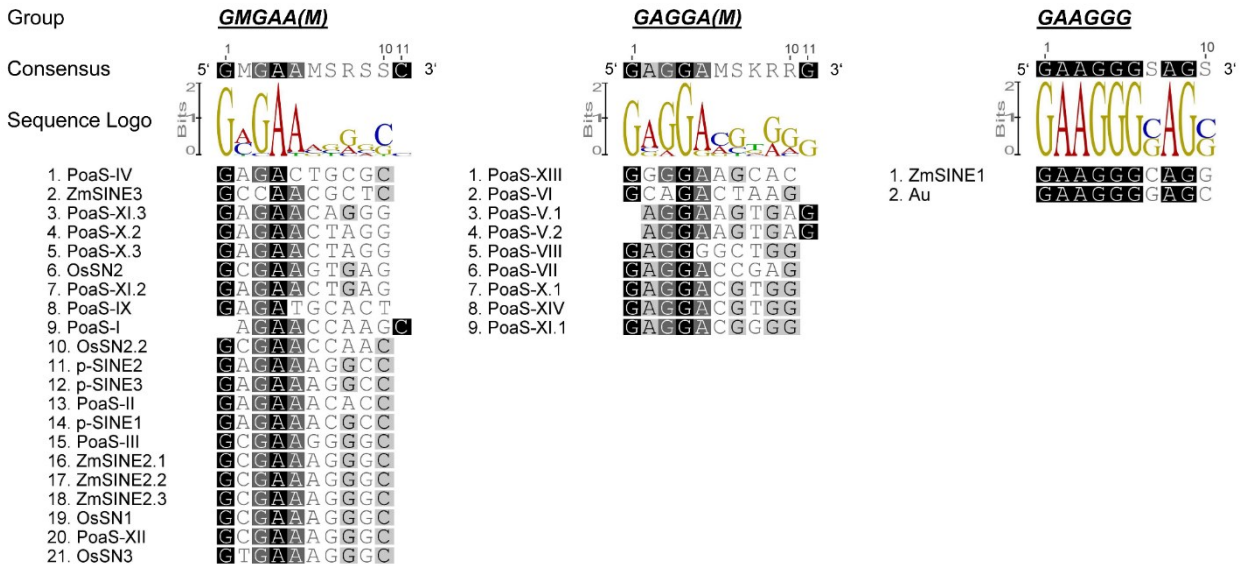
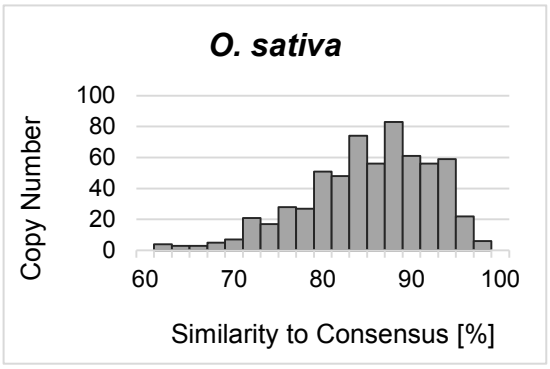
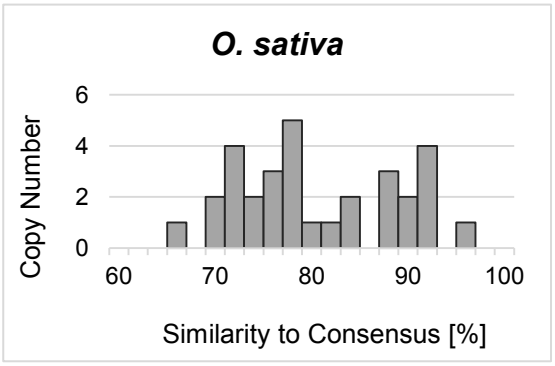


Figure S6. Conservation of 5' start motifs of Poaceae SINE families and subfamilies. PoaS families and subfamilies fall into three different groups concerning the first six nucleotides of the 5' end. For each group, the first ten 5' nucleotides of the consensus sequence of all Poaceae SINE families and subfamilies are shown. The consensus sequence of the start motifs and their respective sequence logo are shown above. Similarity shadings: black – 100 %; dark grey – 80 % to 100 %; pale grey – 60 % to 80 %, white – less than 60 %.

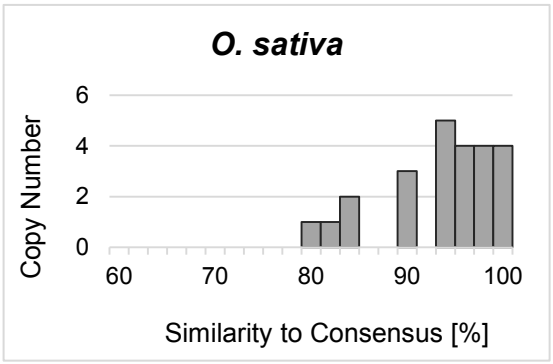
p-SINE1



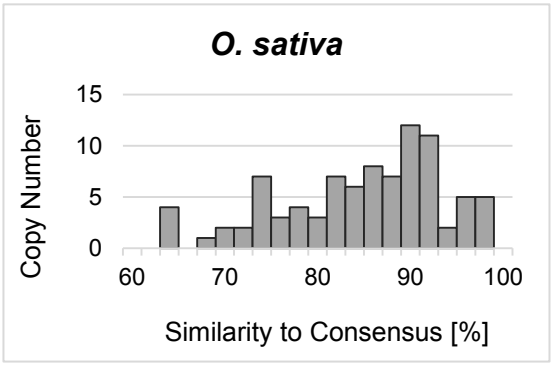
p-SINE2



p-SINE3



OsSN1



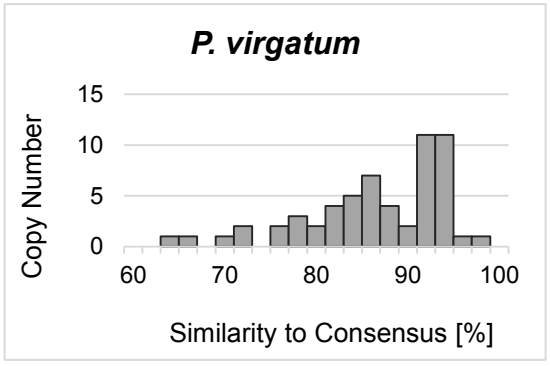
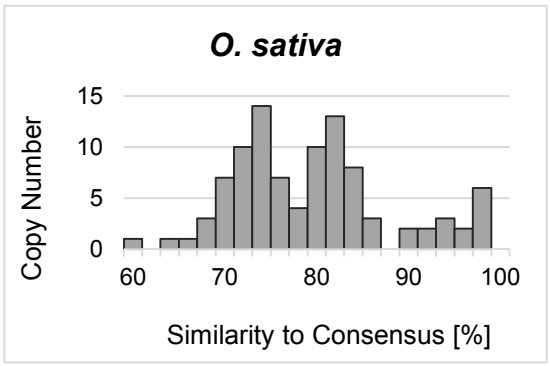
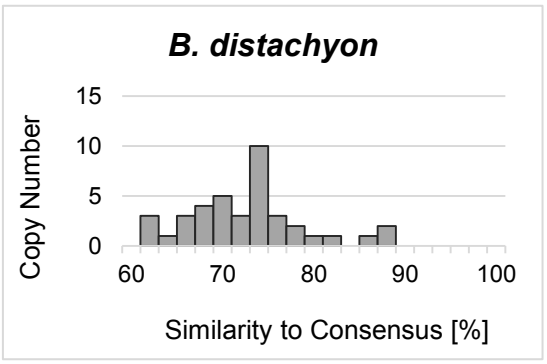


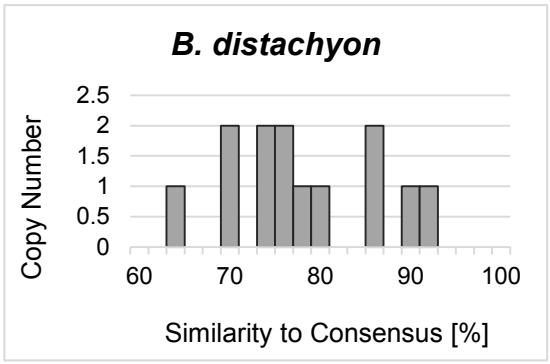
Figure S7. Similarity of SINE family members to their consensus sequence. Poaceae SINE families are represented with the species, for which they were characterized (listed in Table 1). Histograms for other species are supplemented, if the SINE family occurs with at least ten full-length copies.

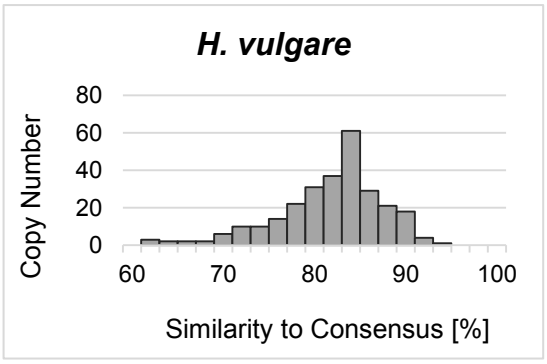
OsSN2.1

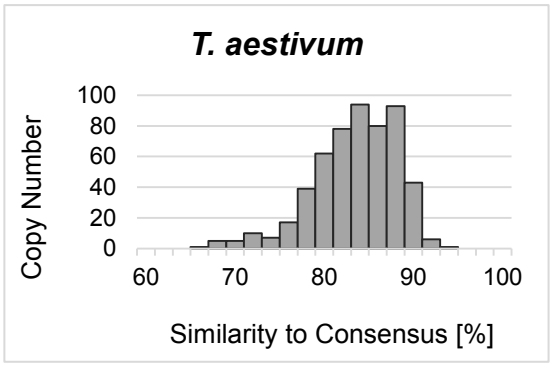


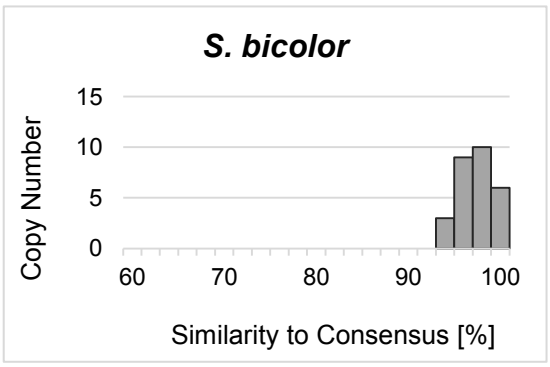


OsSN2.2









OsSN3

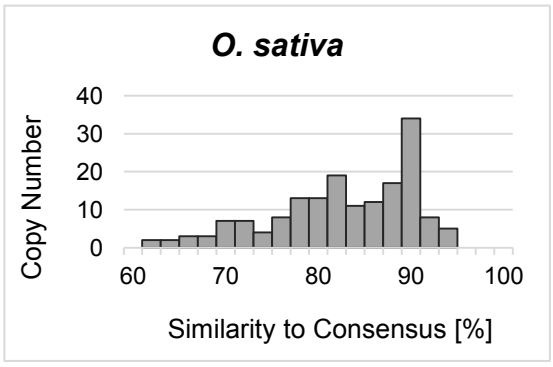
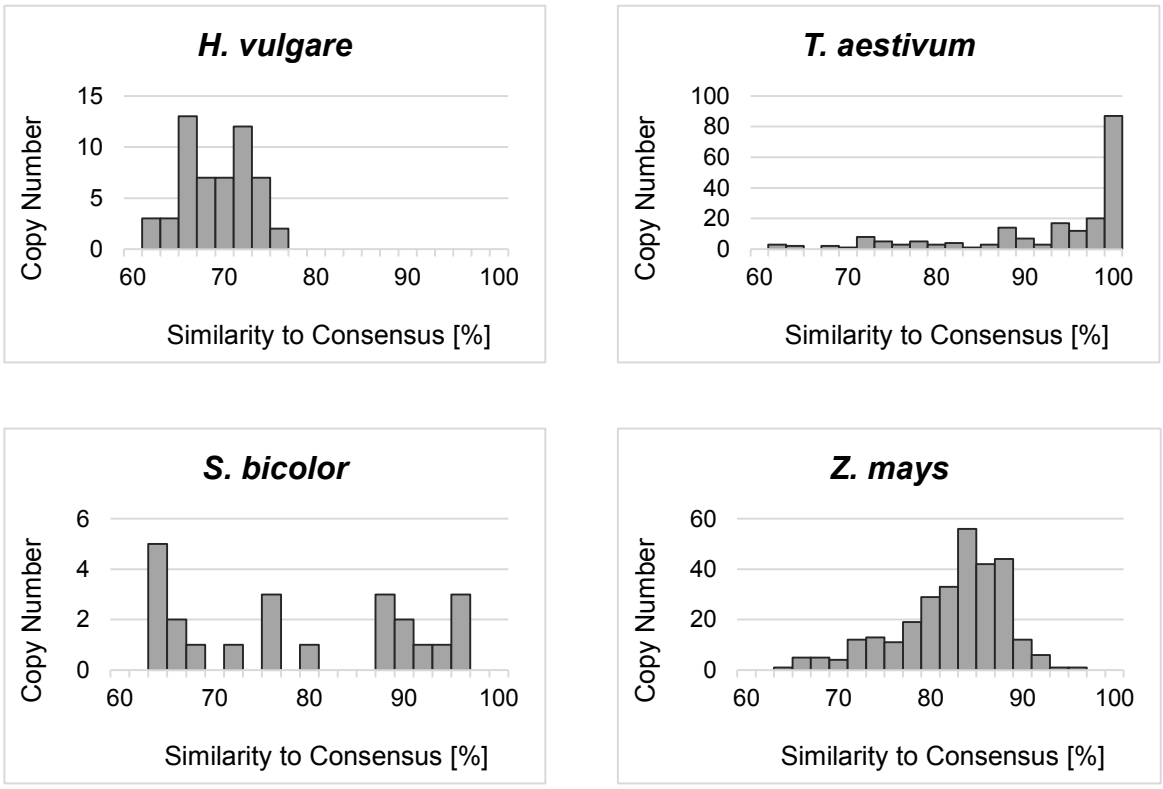


Figure S7. Continued.

ZmSINE1



ZmSINE2.1

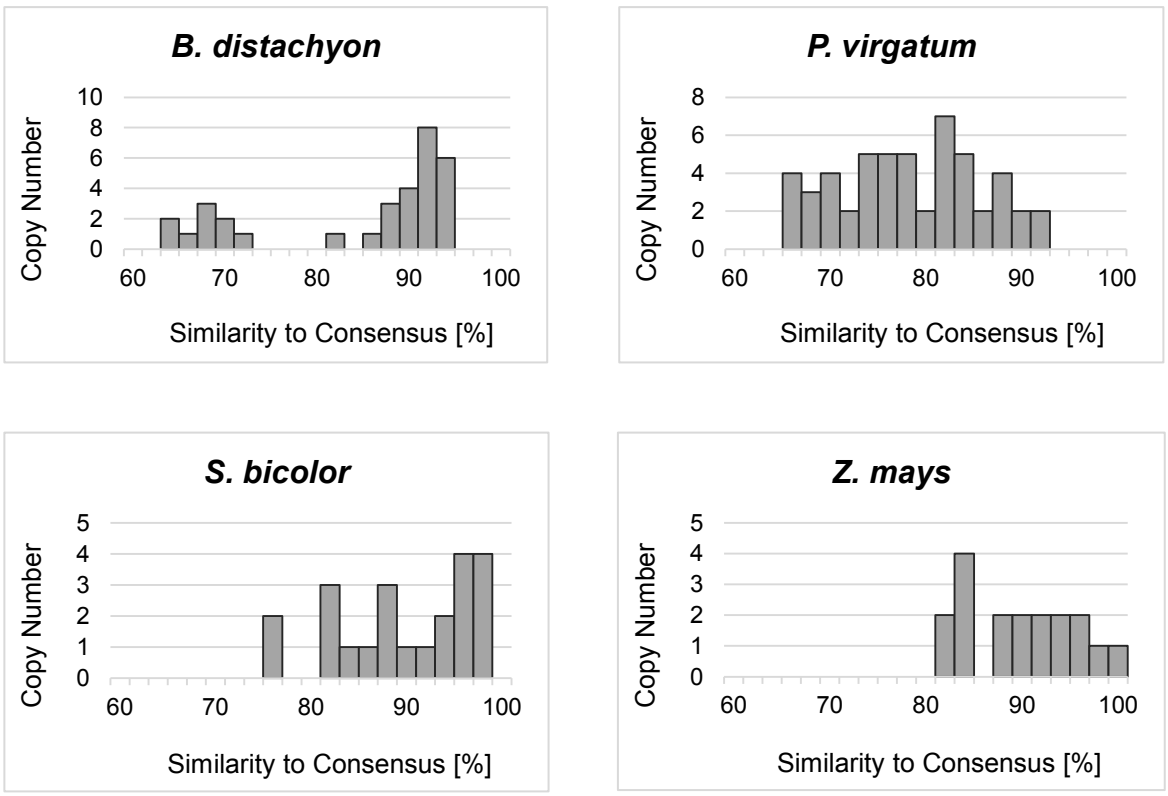
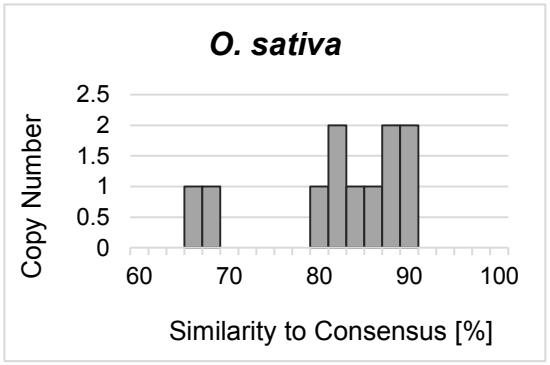
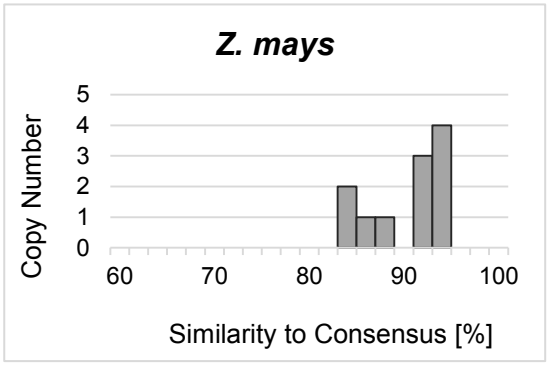


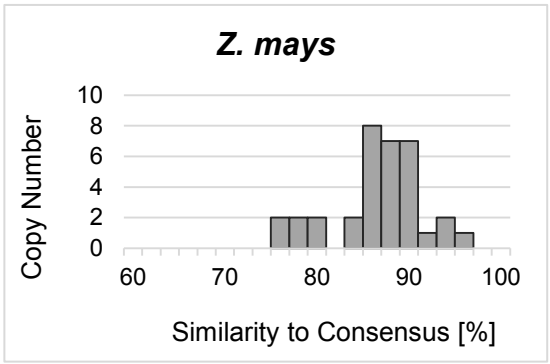
Figure S7. Continued.

ZmSINE2.2

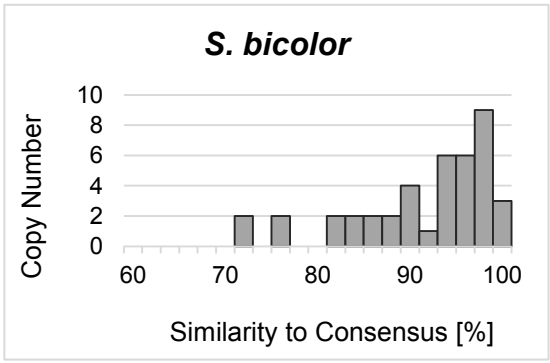




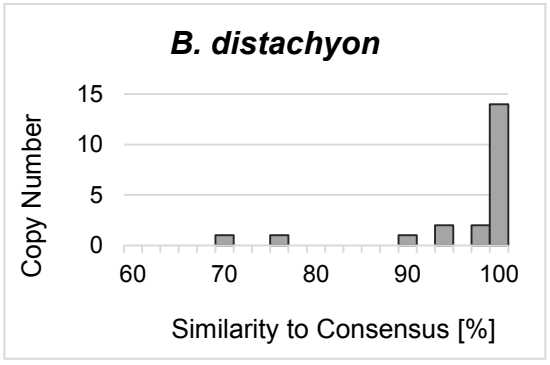
ZmSINE2.3



ZmSINE3



Au



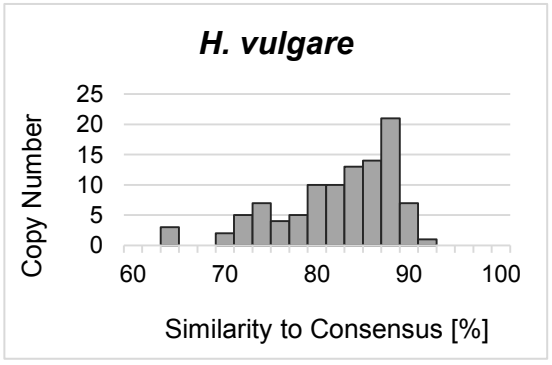
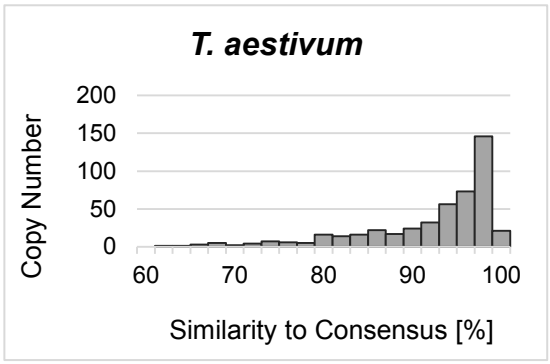
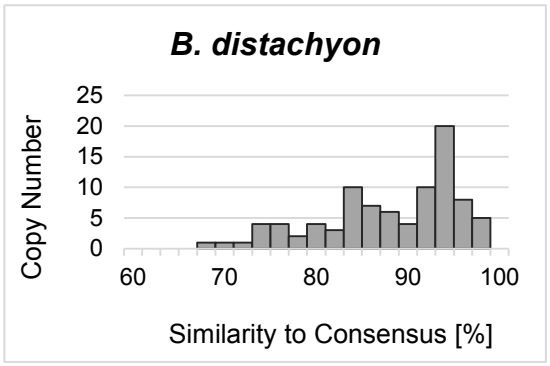


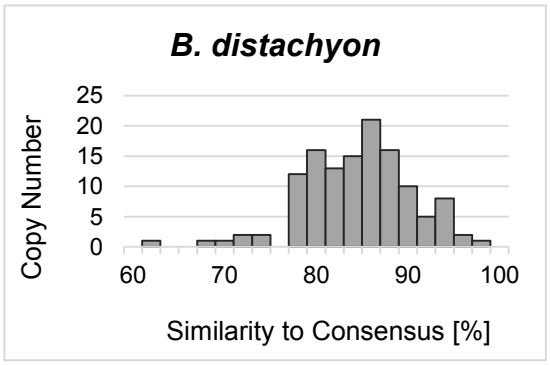
Figure S7. Continued.

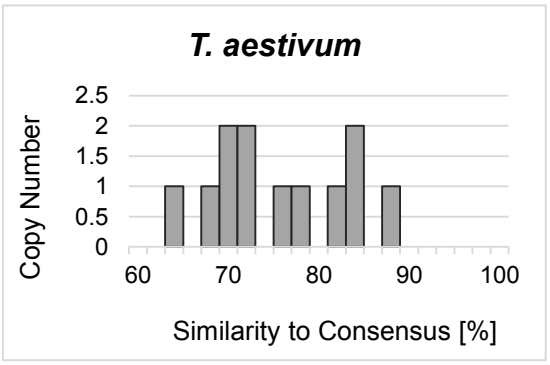


PoaS-I



PoaS-II





PoaS-III

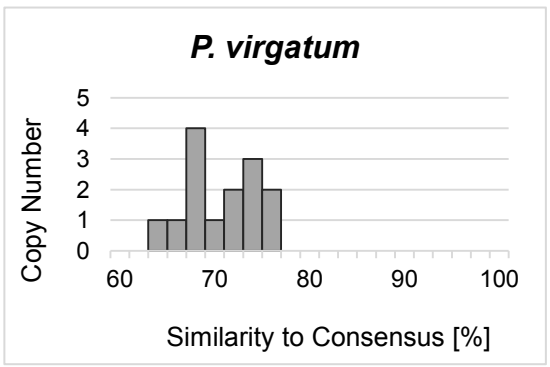
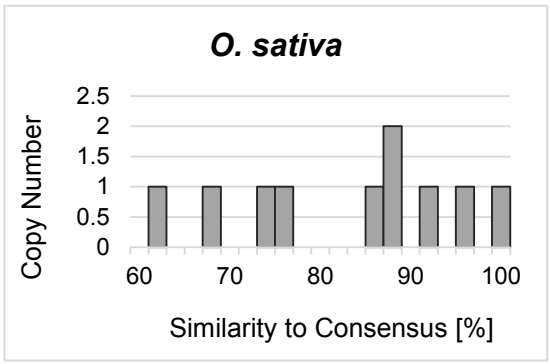
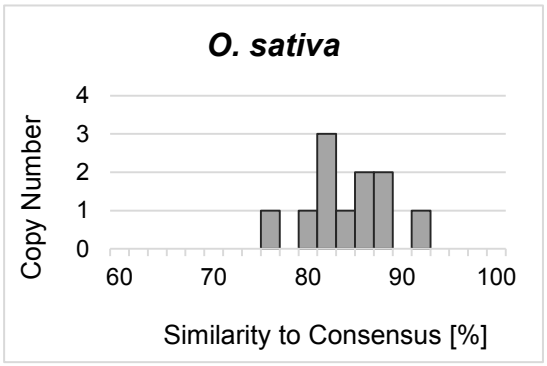


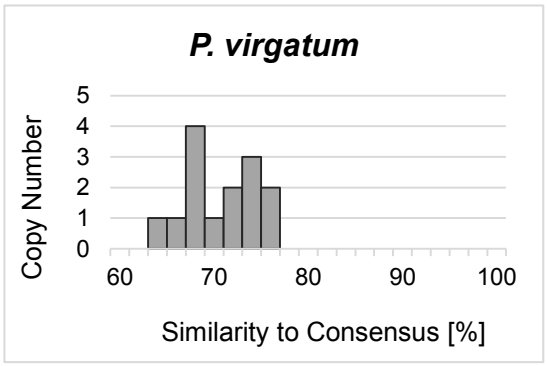
Figure S7. Continued.

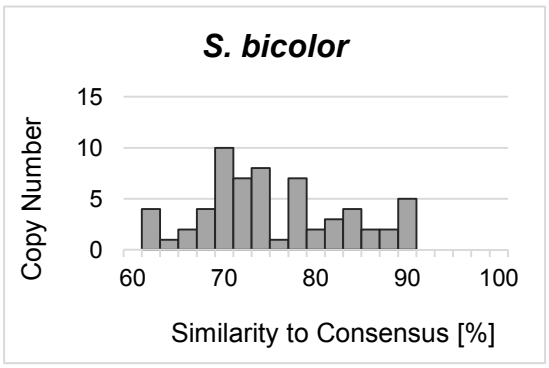
PoaS-IV

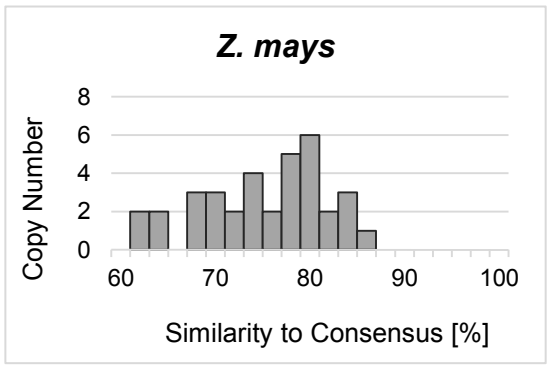


PoaS-V.1

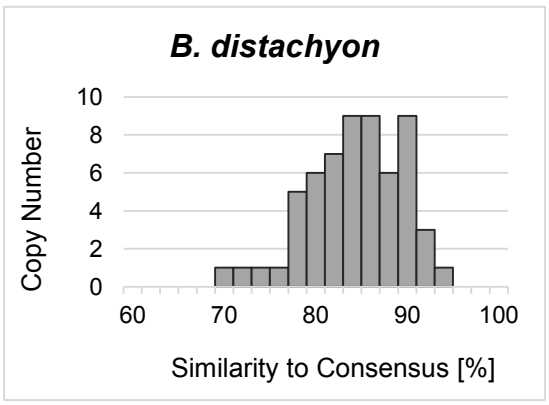








PoaS-V.2



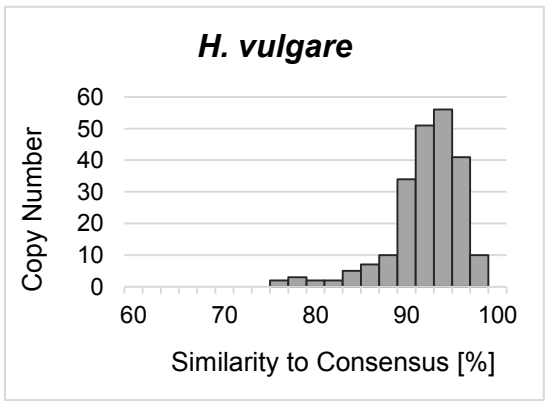
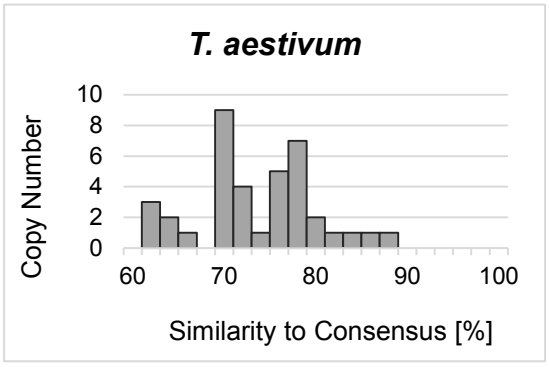
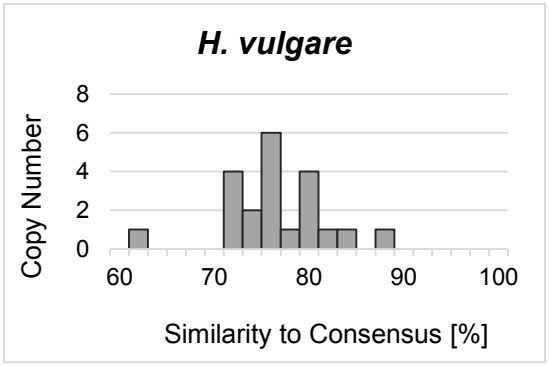
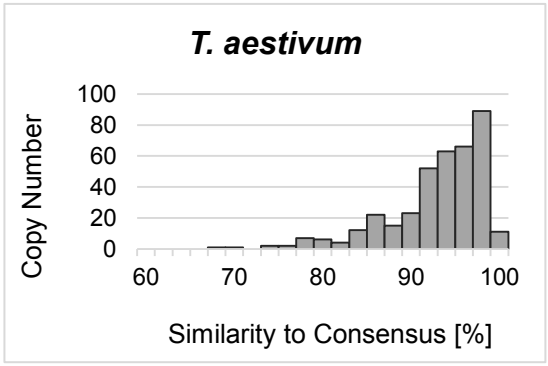


Figure S7. Continued.

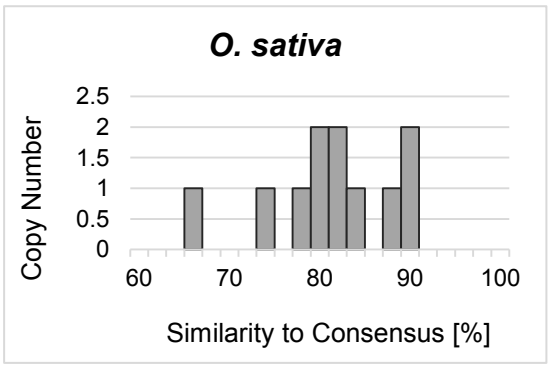


PoaS-VI





PoaS-VII



PoaS-VIII

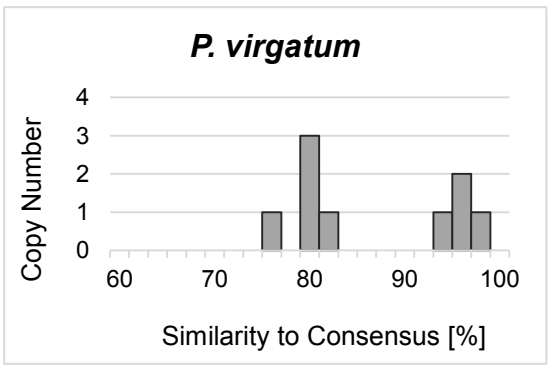
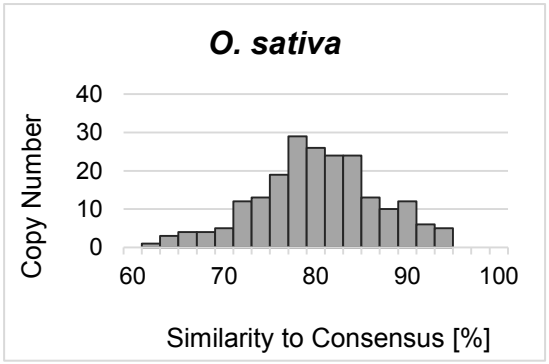
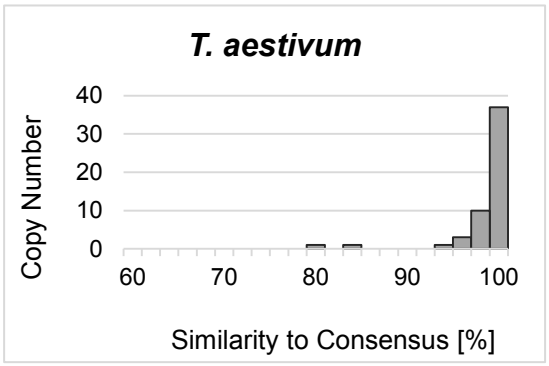


Figure S7. Continued.

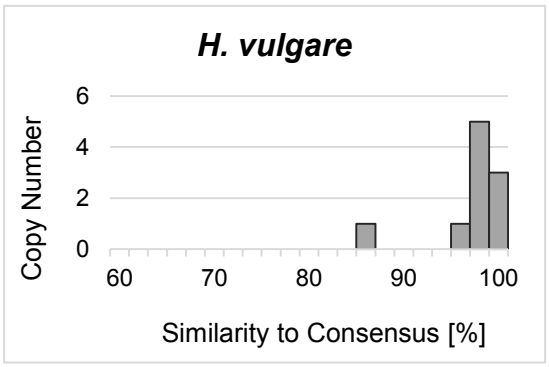
PoaS-IX

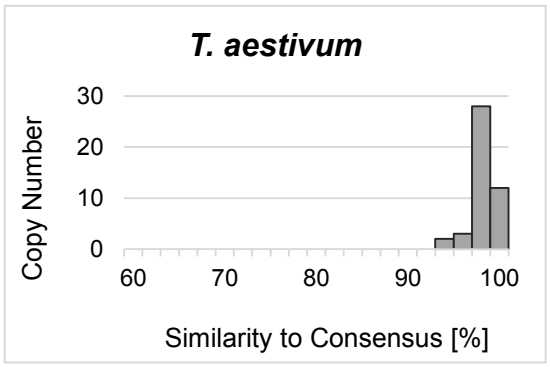


PoaS-X.1

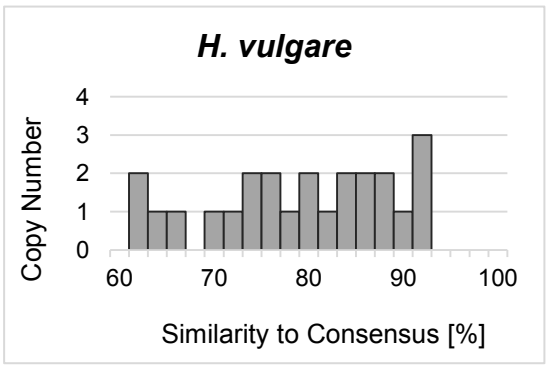


PoaS-X.2





PoaS-X.3



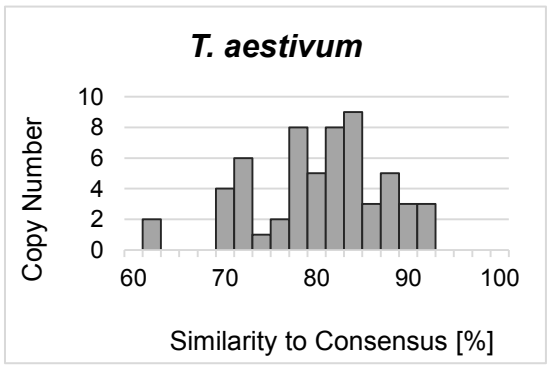
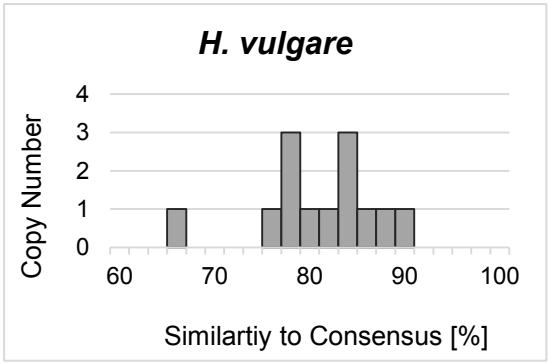
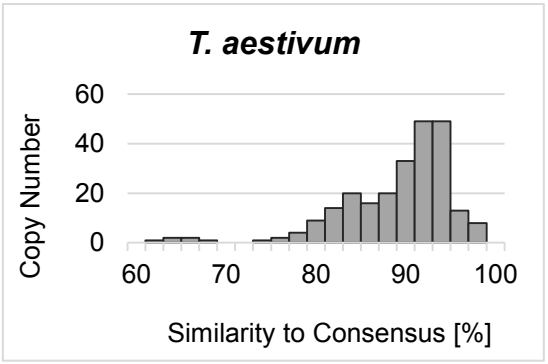


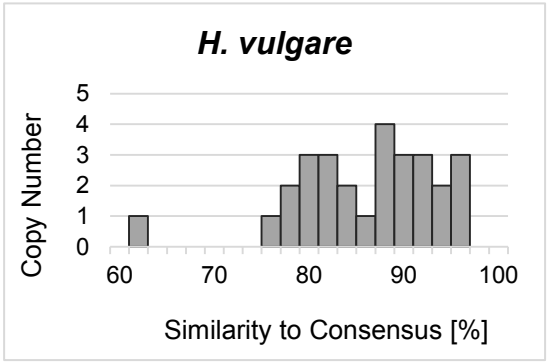
Figure S7. Continued.

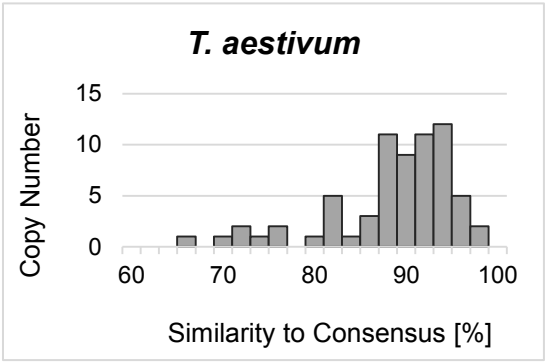
PoaS-XI.1



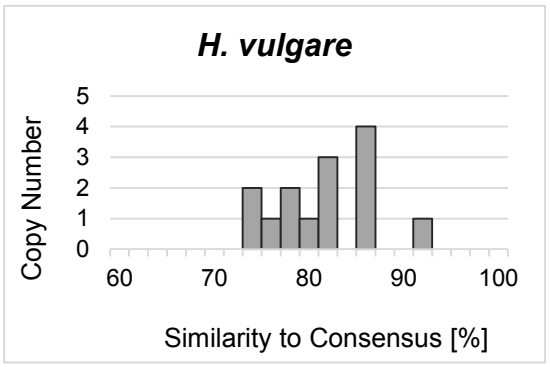


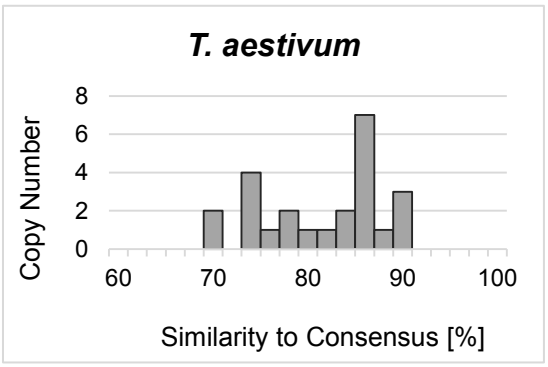
PoaS-XI.2



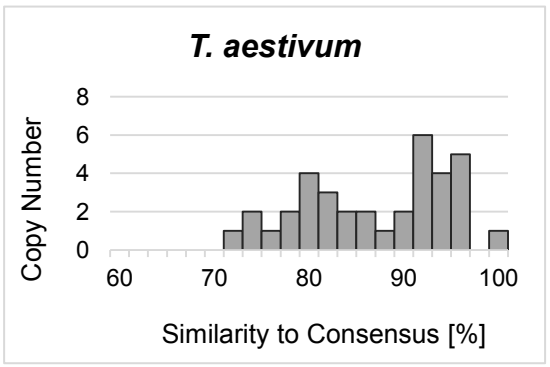


PoaS-XI.3





PoaS-XII



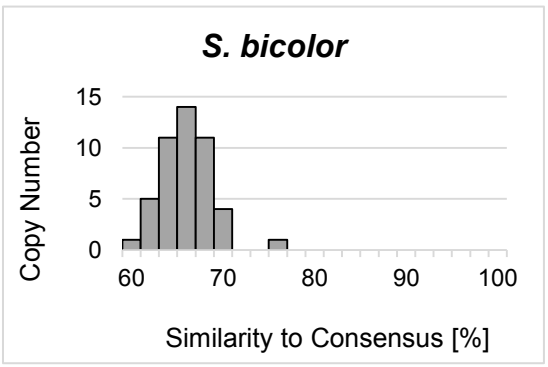


Figure S7. Continued.

PoaS-XIII

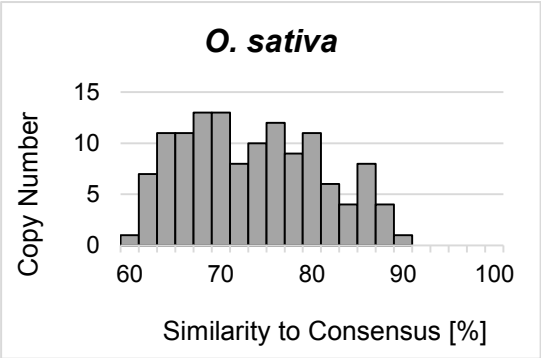


Figure S7. Continued.

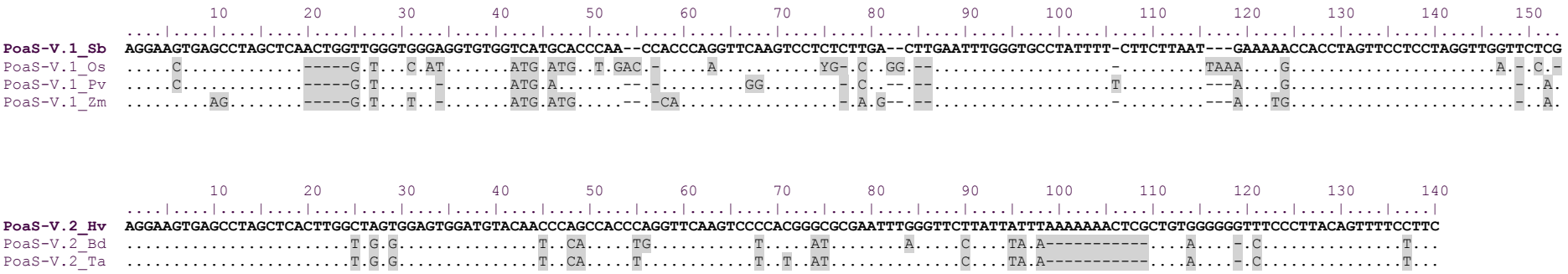


Figure S8. Structural differences between the subfamilies PoaS-V.1 and PoaS-V.2. Species-specific consensus sequences of PoaS-V.1 and PoaS-V.2 were compared. The reference consensus sequences from sorghum millet (PoaS-V.1) and barley (PoaS-V.2) share 65 % sequence similarity. Abbreviations: Bd – *Brachypodium distachyon*, Hv – *Hordeum vulgare*, Os – *Oryza sativa*, Pv – *Panicum virgatum*, Sb – *Sorghum bicolor*, Ta – *Triticum aestivum*, Zm – *Zea mays.*

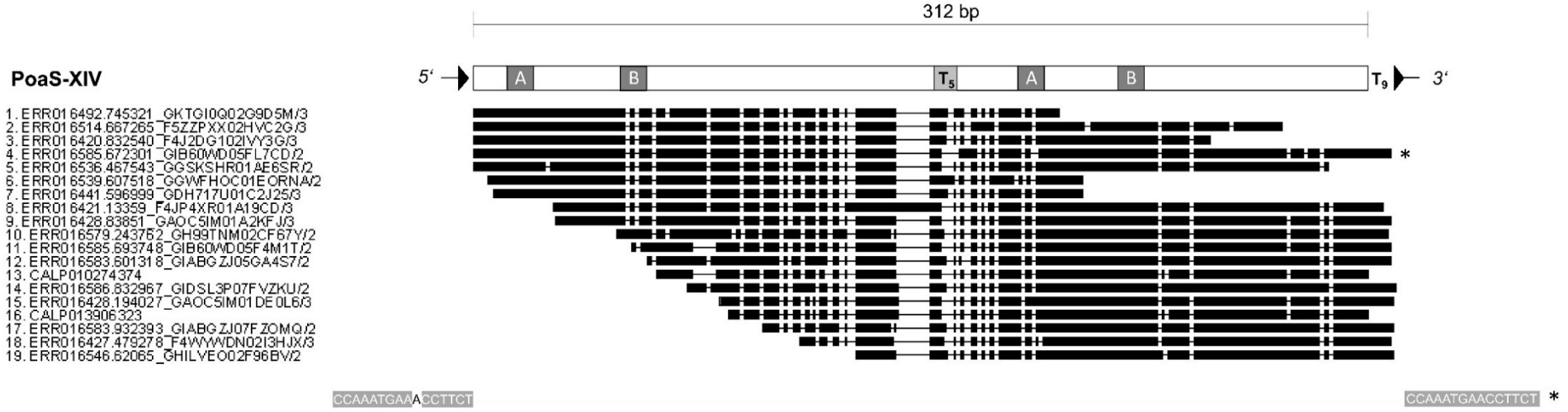


Figure S9. Structure of the homodimeric SINE family PoaS-XIV. Schematic representation of PoaS-XIV (above), originated from two tandemly arranged PoaS-X.1 copies. The white rectangle (top) represents the SINE PoaS-XIV, containing twice the boxes A and B and an internal T-stretch with an average length of 5 bp. The consensus element (312 bp) and the terminal poly(T) tail of 9 bp are flanked by target site duplications (TSDs), indicated as black triangles. The schematic alignment of PoaS-XIV sequences is arranged to the structure above. The only full-length copy found in wheat is marked with by star. The 16 bp TSD of PoaS-XIV is shown below.

Supplemental Tables

Table S1. Genome data sets analyzed in this study.

Analysed species	Sequence data		
	URLs, source	**Size [Mb]**	**Total size [Mb]**
Brachypodium distachyon	http://www.ncbi.nlm.nih.gov/nuccore/?term=Brachypodium+distachyon+[Organism] http://www.ncbi.nlm.nih.gov/nucest?term=Brachypodium%20distachyon%20[Organism] http://www.ncbi.nlm.nih.gov/nucgss?term=Brachypodium%20distachyon%20[Organism]	1.120	**1.481**
	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (addn - release 114)	361	
Oryza sativa japonica	http://www.ncbi.nlm.nih.gov/nuccore/?term=Oryza+sativa+japonica+[Organism] http://www.ncbi.nlm.nih.gov/nucest?term=Oryza%20sativa%20japonica%20[Organism] http://www.ncbi.nlm.nih.gov/nucgss?term=Oryza%20sativa%20japonica%20[Organism]	3.598	**6.552**
	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (aacv, babo, bacj - release 114)	2.954	
Sorghum bicolor	http://www.ncbi.nlm.nih.gov/nuccore?term=Sorghum%20bicolor%20[Organism] http://www.ncbi.nlm.nih.gov/nucest?term=Sorghum%20bicolor%20[Organism] http://www.ncbi.nlm.nih.gov/nucgss?term=Sorghum%20bicolor%20[Organism]	3.210	**4.173**
	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (abxc, ahao, ahap, ahaq - release 114)	963	
Zea mays	http://www.ncbi.nlm.nih.gov/nuccore/?term=Zea+mays+[Organism] http://www.ncbi.nlm.nih.gov/nucest?term=Zea%20mays%20[Organism] http://www.ncbi.nlm.nih.gov/nucgss?term=Zea%20mays%20[Organism]	5.652	**6.281**
	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (ahid, aeco - release 114)	629	
Panicum virgatum	http://www.ncbi.nlm.nih.gov/nuccore?term=panicum%20virgatum%20[Organism] http://www.ncbi.nlm.nih.gov/nucest?term=panicum%20virgatum%20[Organism] http://www.ncbi.nlm.nih.gov/nucgss?term=panicum%20virgatum%20[Organism]	1	**1**
Hordeum vulgare	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (cajw - release 115)	1.868	**1.868**
Triticum aestivum	*http://www.ebi.ac.uk/ena/data/view/ERP000319*	128.000	**132.237**
	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/wgs/ (calp, calo - release 115)	4.237	

Table S2. Consensus sequences of Poaceae SINE families. Consensus sequences of previously published SINE families used for initial *BLAST* searches were obtained from databases or corresponding publications.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
AU	*Aegilops umbellulata*	GAAGGGGAGCCTTGGCGCAGTGGTAAAGCTGCTGCCTTGT GACCATGAGGTCACGGGTTCAAGTCCTGGAAACAGCCTCT TACAGAAATGTAGGGAAAGGCTGCGTACTATAGACCCAAA GTGGTCGGACCCTTCCCCGGACCCTGCGCAAGCGGGAGCT ACATGCACCGGGCTGCCC	178	poly(T)	Yagi *et al.*, 2011; this study	n.a.
OsSN1	*Oryza sativa*	GCGAAAGGGCCTGTAGCCTAGTGGTTACAAGAGCCTCAGT AGCACCTGAGGTCCTGGGTTCGACTCCCCATGGGAGCGAA TTTTCCAGGATTTAACGGCGTTGTGCTTTCAGTGGTAGGCG ACGTACCCGTCGACAGCGAGGCGCCTGTGGTGACTTCGTC AATCTCTCAGGATTTGCCGGCCCAGTCTTCGAAGATGCTCA TAGGGGTAGGGTTTGCGTGCGTGCGTTCATAGGGGTGAGT GTGCGTGCGTTGTGAGTGTCTGCGTTGTACTGTGTAATTCT	283	poly(A)	Tsuchimoto *et al.*, 2008; this study	AB427154
OsSN2.1	*Oryza sativa*	GCGAAGTGAGCGTAGCTCAACTGGTTAGGTTCCTTGTGGT GGAACCAGCCCACCCGGGTTCAAATCCTAGATTTGACACG GGTGCTCGCATTTACGGCTAATTATTCTTTCAGTGGTAGGCG ACGTACCCGTCGACAGCGAGGCGCCTGTGGTGACTTCGTC AATCTCAAGATATGTCGGCCCAGTCTTTCGGAGGTGCTCAT AGGGGTAGGGTGTGCGTGTGTGCGTTCATAGGGGTGAGTG TGCGCGCGTTGTGAGCGCCTGCGTTTGTACTGTGTTTCT	282	poly(A)	Tsuchimoto *et al.*, 2008; this study	AB427155
OsSN2.2	*Triticum aestivum*	GCGAACCAACCTGTGGTTGGATGGTTAGAGGGACAGTGGT ATCCCCAGCCCACCAGGGTTCAAGTCCTTATTCCTGGATTT ATTTCAGGATTTCCGGCGATGCGCATTCAGTGGGAGGAGA CGTTCCCGTCGACGACGAGGCGCCTACGGTGACTTCGTAA ATTTCAAGATGATATGCCGGCTCAGTCTTTCGGAGGTGCTC ATAGGGGTAGGGTGTGCGTGTGTGCGTTCATAGGGGTGAG TGTATGCGCGTGTATATGAGCGCTTGCGTCTGTACTGTGTT	283	poly(A)	this study	n.a.
OsSN3	*Oryza sativa*	GTGAAAGGGCATGTAGCCTAGTGGTTGCAGTGACCTGAGT AGCACCCCAAGGTCCTGAGTTCAAATCTCCATAGGAGCGA ATTTCAGATTGGGTTGTTTGAGGGGCTAAGTTCCCAATTTA AATGGCTGCATATATCCGGTTGGATGTAGAGGCCGGGTAAA AAATACCCTTCTCT	176	poly(A)	Tsuchimoto *et al.*, 2008; this study	AB427156

[a] where the SINE family was identified

[b] for Consensus sequence

n.a. not available

Table S2. Continued.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
p-SINE 1	*Oryza sativa*	GAGAAACGCCCAGGGGTCTTCCGGCTAGCTCCACAAGGT GGTGGGCTAGACGACCTGGGTTCGAAGCCTCACCCCTTCT AATTATTTGATATTAGGTCATTCCCTAATATTCGCG	115	poly(T)	Mochizuki *et al.*, 1992; this study	n.a.
p-SINE 2	*Oryza sativa*	GAGAAAGGCCCRGGGGTCTTCCGGCTAGCACCACAAGGT GTGGGCTAGCCGACCTGGGTTCGAGCCTCACCCCTCTTAA TAAATTTCGATATGAGAGCCCCTCCTCTCATATCCAGCG	118	poly(T)	Xu *et al.*, 2005; this study	AB206875
p-SINE 3	*Oryza sativa*	GAGAAAGGCCCGGGGTCTTCCCGGCTAGCAACGCAAGCT GCGAGCTAGCCGGTCCGGGTTCGAGCCTCACCCCCTCCTT AATTCAAAATCAATCTAGTCCTTCCTAGATTGGTCCCA	117	poly(T)	Xu *et al.*, 2005; this study	AB206894
PoaS-I	*Brachypodium distachyon*	GAGAACCAAGCATAGCTTGGGTGGTCAGCCAGCCAGGTAT GCTGGCAGCCCACCAGAGTTCGATCCTCGAAGGTCGCACT TTGGTGTCTCACTTTGTAAAAATTATATATCTATATACTGTC GGACTGCAATCGCGCAGCTCTTACAGTTTAAATTC	157	poly(A)	Wenke *et al.*, 2011; this study	n.a.
PoaS-II	*Brachypodium distachyon*	GAGAAACACCTCTTGGTGTGGTGGTGGAGTTGTGGGTGC ATGACTCCACCCACCAGGGTTCAAATCCTGGTGCTCACAA TTATGCTTTAGGGGTTTCCCTTACAGTCTTTCCAT	114	poly(A)	Wenke *et al.*, 2011; this study	n.a.
PoaS-III	*Panicum virgatum*	GCGAAGGGGCTCACGGTGCAGTGGCAAAGGCCACTGGTC GGGGTGCTCCCGCCCAGGGTTCAAACCCTGGGTGCCGCA CCTTTAAGCTTCAGGGGTTTCCCTTRGAGTATTTCTATC	117	poly(A)	this study	n.a.
PoaS-IV	*Oryza sativa*	AGGAACTGAGCCTAGCTCAGTTGGTCGATGGTGTGGATGT ATGCCTAGACCACCCAAGTTCAAGTCCTYGTCGAGGCGAA TTTGGGTGCCTATTTTCTTCTTAATACAAAAGCCACCTAGT TCCTCCTAGGTTGATCCC	139	poly(T)	this study	n.a.
PoaS-V.1	*Sorghum bicolor*	AGGAAGTGAGCCTAGCTCAACTGGTTGGGTGGGAGGTGT GGTCATGCACCCAACCACCCAGGTTCAAGTCCTCTCTTGA CTTGAATTTGGGTGCCTATTTTCTTCTTAATGAAAAACCAC CTAGTTCCTCCTAGGTTGGTTCTCG	145	poly(T)	this study	n.a.

[a] where the SINE family was identified

[b] for Consensus sequence

n.a. not available

Table S2. Continued.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
PoaS-V.2	*Hordeum vulgare*	AGGAAGTGAGCCTAGCTCACTTGGCTAGTGGAGTGGATGTACAACCCAGCCACCCAGGTTCAAGTCCCCACGGGCGCGAATTTGGGTTCTTATTATTTAAAAAAACTCGCTGTGGGGGGTTTCCCTTACAGTTTTCCTTC	140	poly(A)	this study	n.a.
PoaS-VI	*Triticum aestivum*	GCAGACTAAGGCATAGCCTAGTGGTGGGAAGGGGCTGATGCCTTCCCACCCACCCAGGTTCAAGGCATGGTACTTGCAATTTGGGTTTGTTGCACCAATTATACTGTAGGGGGTTCCCTTACAGTCTTTCTGTC	134	poly(A)	this study	n.a.
PoaS-VII	*Oryza sativa*	GAGGACCGAGCGTTGCTCGGGTGGCAAGCGYCGCTGGTGCGCKCGCTGCCCACGAGCGTTCGAWCCCTGGGATCGCAACTCTCGTGCCTCCCGGGGGGGATTTTTCCCCTCTTTCCCCGGGACTGACTTCGGTTGGTCCCGGCTAGGTAATAGGGTACACACACGCGTGCGCGTTCAGTGGGACTGCACGTTTCCCGTGCACTGAGGCCTAGTGCCTCCAATCTCATTCCTAGCGTGTGTTAGGGACGCGCACGCGTGTGTGTGCGTGTGTYGTGGTGTGAGTGTGGTGTGTGTAAGTGTGCGTCCTGAGTTGTACCCTTCT	321	poly(A)	this study	n.a.
PoaS-VIII	*Panicum virgatum*	GAGGGGCTGGTGAGCCCGAGTGGCTCCTGGAGCCAGCCCCCAGGCCGGCGACCGGGGGTTCGATCCCCCGSGCTGGCACCGGGGAGGCCCTCTGTTACCTCTCCTAGTG	108	poly(A)	this study	n.a.
PoaS-IX	*Oryza sativa*	GAGATGCACTTGATAGTGCAGTGGCAAGGGGTGTGTGGTTTCAACCCTGAGGTCCCGTGTTCAATCCCCAACACGCTCATAATTTCTTCTTAAAATGTTTGGAGGGACGTCTCTCCCTCCAAATCTCG	128	poly(T)	this study	n.a.
PoaS-X.1	*Triticum aestivum*	GAGGACGTGGGCATAGCCCAGTGGTTGGGGGCGCATGATTGTAAACCTAACGACCAGAGTTCGATCCACGTCGGGGACGAATTTCTGGAATTCTCATGAGGGATGCTTCTTCTATATCAATAAAACCGTGGGTGCTAGTGCCCATGGAGTTTCA	154	poly(T)	this study	n.a.
PoaS-X.2	*Triticum aestivum*	GAGAACTAGGCTGTAGCCTAGTGGCAAGGGAGCGCAGTGGCGTCTCCAGCAACCAGGGTTCGAGCCACGTCGGGGACGAATTTCTGGTTTCTCACAAGGGATGCTTCTCCTATATCAATAAACCATGGGTGCTAGTGCCCATGAGTTTCATC	152	poly(T)	this study	n.a.

[a] where the SINE family was identified
[b] for Consensus sequence
n.a. not available

Table S2. Continued.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
PoaS-X.3	*Triticum aestivum*	GAGAACTAGGCTGTAGCCTAGTGGCAAGGGTCGCAGTGGCGCACCCTGCGGCCAGGGTTCGACTCCCGTCGGGAGCGAATTTCTGGTACCTCATCCGGGTGGGCTTCTTCTATAAAAATATGTCCTGGGTGCTAGTGCCCATGGATCTCA	150	poly(T)	this study	n.a.
PoaS-XI.1	*Triticum aestivum*	GAGGACGGGGCGTCGACCCGTTGGCTGGGCAGCTGAGGTTGCTGCCAGCCCACCCGAGTTCGAGTCCCGGCTCGGACGCGCGGTGCTCGCGGAGTTTCTCCTATAAAAAAATGCCAACGAGGGTTAGCCCTTGGGTTGGTCTCA	144	poly(T)	this study	n.a.
PoaS-XI.2	*Triticum aestivum*	GAGAACTGAGCGTAGCTCAGTTGGCAAGGCGCGGGAGTTCGCAGCCAGCCCACCAGGGTTCAAGTCTCGGCTTGAGCGTTTGGTGCTCACGGAGTTTTCTTCTATAAAAAAATGCCAACAGGCTAGTCTAGCCCGGGTTGGTCTCG	146	poly(T)	this study	n.a.
PoaS-XI.3	*Triticum aestivum*	GAGAACAGGGTGTCACCCTGTTGGCTAGGCTHGCGCGGAGCCAGCTAGCCCACCCGGGTTCGAGTCCCGGAGTGGCCCCGTGGTGCTTAGAGATTTCTTCTATAAAAATATGCCTCCAAGGGCTAGTCCTGGATGGTCTCG	141	poly(T)	this study	n.a.
PoaS-XII	*Sorghum bicolor*	GCGAAAGGGCCTCTAGCCTAGTGGTTAGAGCACCTGAGTAGCACCAGCAGACCTGGGTTCGACTCCCCGTGGGAGCGAATTTAAACAGGTCTGCATTAAAAAAAAATAAAAAATAGGCTGGGGTTTCCCTTGCTGACTTCGGTC	144	poly(A)	this study	n.a.
PoaS-XIII	*Oryza sativa*	GGGGAAGCACCAGTGGTGTGGTGGTGGAGTCGTGGGTGCATGACTCCACCCACCAGGGTTTAAATCCTGGTGCCCACGAATATTACGCACATGTAGGTGGACTTTCAATAGGATTTTAGTGAGATCAGGGATGTGCCGCTGGTTTCCGTCTCTTAGAGCATGTGTTAGGGGACGCATTCGTGGGGGTGTGAGTGTGGTGTTGCGTGTGTAGTGGTGTGTGTGCGTGTGCGTCTGCCGTGTAATT	244	poly(A)	this study	n.a.

[a] where the SINE family was identified

[b] for Consensus sequence

n.a. not available

Table S2. Continued.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
PoaS-XIV	*Triticum aestivum*	GAGGACGTGGGCATAGCCCAGTGGTTGGGGGCGCATGATT GTAAACCCAACGACCAGAGTTCGATCCACGTCGGGGACG AATTTCTGGAATTCTCATGAGGGATGCTTCTTCTATATCAAT AAACCGTGGGTGCTAGTGCCCATGGAGTTTCATTTTTNGA GGACGTGGGCATAGCCCAGTGGTTGGGGGCGCATGATTGT AAACCYAACGACCAGAGTTCGATCCACGTCGGGGACGAA TTTCTGGAATTCTCATGAGGGATGCTTCTTCTATATCAATAA AACCGTGGGTGCTAGTGCCCATGGAGTTTCA	312	poly(T)	this study	n.a.
ZmSINE1	*Zea mays*	GAAGGGCAGGCCTGGTGCAGTGGTGAGAGCTGTCTCACT GAGTCACCAGGTCGCGGGTTCGAAGCAGCCTCTCCGCATT TGCGGGGGAAGGCTTGCCTCGGTTTATCCCTTCCCCAGAC CCCACTCATGTGGGAGCCTCCGGCACTGGGTCTGCCC	156	poly(T)	Baucom *et al.*, 2009; this study	RST_ZmSINE1_ consensus-0
ZmSINE2.1	*Panicum virgatum*	GCCAAAGGGTGTCTAGCCGGATTGGTTAGGTGGCCCCAGC GGCACTCCTCAGGTCCTGGGTTCGACTCCCGGTGGGAGCG AATTTCAGGCTGAGGTTAAAAAAATCCCCTCGCCTGCCTC ATGTCCAAAGCACTGTGGAGCCCGGCCTAACTCACAAGG CGACGGGCCCCCGTGTACGGGTGGGGCAGGGGTTCGGGG GTTTTCTTGGCCTGCTGTGAGAGGTCATTCTACCTCTCAAA CAATGCCGTGGGGGCGGCTTACCCCCCGCAGGTCAAG	276	poly(T)	Baucom *et al.*, 2009; this study	RST_ZmSINE2.1_ consensus-0
ZmSINE2.2	*Zea mays*	GCGAAAGGGCCTCTAGCTGAGTTGGTTAGGTGGTCTGAGT AGCACTCCTTAGGTCCTGAGTTCGAATCCCAGTGGGAGCG AATTTCAGGCTGAGGTTAAAAAAGGTCACTCGCTGGTTCC CCTGGTTGTGTGCACACGAGATGGACTGACCTATGGGGGG CGGATCCTCGTGTAGGGGCTGGGAGGGCTCAAAGCACGA GTAAAGATCTGGCCTATAGGGGGCGGACCCTCATGTTGCA CGGGGGACCAGCTTTCGTGACCTTTCTCGGTCGGGGCTCC GATTGAGCTTCTTAATATAATACCGTGGGGGCGGTCTTTCC CCTACCGGCCGAG	333	poly(T)	Baucom *et al.*, 2009; this study	RST_ZmSINE2.2_ consensus-0

[a] where the SINE family was identified

[b] for Consensus sequence

n.a. not available

Table S2. Continued.

SINE family	Species[a]	Sequence		Poly(A/T)	Reference[b]	Accession
		Consensus (5' → 3')	Length [bp]			
ZmSINE2.3	*Zea mays*	GCGAAAGGGCCTCTAGCGTAATGGTTAAGGCTTCCGAGTA GCACCTCCAGGTCCYGGGTTCGATCCCCCTCGGGGGCGAA TTTCGGGCTTGGTTAAAAAAATCCCCTCGTTGTGCCCCATC CGCTCTCGGGTTNGATGTCCTGCGCGCCACCCTCCGGYTG GGCCGTTGCAGAGTGGACGGTTGGCCGGCCCGTTAGTGAT GGGGGGCCAGGGTTCGGGGATTTTCTCGGCCGGGACCAT GTTTCGGTCTCTTCTTAATATAATACCGGGAGGGCGGTCTT TCCCTCCCCGGCCGAG	297	poly(T)	Baucom *et al.*, 2009; this study	RST_ZmSINE2.3_ consensus-0
ZmSINE3	*Zea mays*	GCCAACACTCTCACGGTGTAACTGGTCAGCACAACACGC CAAAGAAGCGGTTGGCTGAGCCAGCCCGGGTTCGAGTCA CGGCACCATCTTCTTAAGACGAAAATCAGGGGGACGTCTC TCCCCCTGGTCGAG	132	poly(A)	Baucom *et al.*, 2009; this study	RST_ZmSINE3_ consensus-0

[a] where the SINE family was identified

[b] for Consensus sequence

n.a. not available

Table S3. Distribution of Poaceae SINE families in seven Poaceae species. The total copy number (full-length and 5' truncated) of all Poaceae SINE families and subfamilies are listed per species.

SINE family	Species							Total[a]
	O. sativa	*B. distachyon*	*H. vulgare*	*T. aestivum*	*P. virgatum*	*S. bicolor*	*Z. mays*	
p-SINE1	758	7	-	-	-	-	-	**765**
p-SINE2	43	-	-	-	-	-	-	**43**
p-SINE3	27	-	-	2	-	-	-	**29**
OsSN1	411	-	-	-	151	57	-	**619**
OsSN2.1	255	295	-	-	-	-	-	**550**
OsSN2.2	-	208	1,155	1,250	-	72	-	**2,685**
OsSN3	391	-	-	-	-	-	-	**391**
ZmSINE1	-	12	90	294	203	38	474	**1,111**
ZmSINE2.1	5	56	-	-	88	66	23	**238**
ZmSINE2.2	23	-	-	-	-	-	43	**66**
ZmSINE2.3	-	-	-	-	-	-	122	**122**
ZmSINE3	-	-	-	-	10	60	2	**72**
Au	-	39	286	1,006	10	3	6	**1,350**
PoaS-I	-	121	-	-	-	-	-	**121**
PoaS-II	-	143	6	13	-	-	-	**162**
PoaS-III	-	14	-	-	36	-	-	**50**
PoaS-IV	11	-	-	-	-	-	-	**11**
PoaS-V.1	13	-	-	-	27	70	66	**176**
PoaS-V.2	-	81	307	49	-	-	-	**437**
PoaS-VI	-	-	25	454	-	-	-	**479**
PoaS-VII	48	-	-	-	-	-	-	**48**
PoaS-VIII	-	-	-	-	12	3	-	**15**
PoaS-IX	305	-	-	-	-	-	-	**305**
PoaS-X.1	-	-	5	75	-	-	-	**80**
PoaS-X.2	-	-	35	53	-	-	-	**88**
PoaS-X.3	-	-	42	104	-	-	-	**146**
PoaS-XI.1	-	-	16	315	-	-	-	**331**
PoaS-XI.2	-	-	47	84	-	-	-	**131**
PoaS-XI.3	-	-	19	31	-	-	-	**50**
PoaS-XII	-	-	5	48	-	51	-	**104**
PoaS-XIII	266	-	-	-	-	-	-	**266**
PoaS-XIV	-	-	-	11	-	-	-	**11**
Total[b]	**2,556**	**976**	**2,038**	**3,789**	**537**	**420**	**736**	**11,052**

[a] SINE copy number per SINE family

[b] SINE number per species

Table S4. Primers used for synthesis of Poaceae SINE probes for fluorescent *in situ* hybridization.

SINE family	Primer	Amplicon [bp]	Nucleotide position [bp][a]	Identity [%][a]
PoaS-X.2	*for* CGTCGGGGACGAATTTCTGG	83	68 - 150	92.8
	rev TGAAACTCATGGGCACTAGC			
ZmSINE1	*for* GGTCGCGGGTTCGAAGCAGC	104	49 - 151	92.3
	rev GGAGCCTCCGGCACTGGGTC			

[a] regarding consensus sequence

Table S5. Intervals of average similarity values of Poaceae SINE families. The average similarity values of 31 Poaceae SINE families and subfamilies were grouped in six similarity intervals (x) each. PoaS-XIV (one full-length copy only) is not included.

Similarity range [%]	Number[a]	SINE families
$x \geq 90$	3	p-SINE3, PoaS-X.1, PoaS-X.2
$85 \leq x < 90$	7	Au, OsSN2.2, PoaS-V.2, PoaS-VI, PoaS-XII, ZmSINE2.2, ZmSINE3
$80 \leq x < 85$	6	OsSN1, PoaS-I, PoaS-VIII, PoaS-XI.1, PoaS-XI.2, ZmSINE2.3
$75 \leq x < 80$	7	OsSN3, p-SINE1, PoaS-II, PoaS-III, PoaS-IV, PoaS-VII, ZmSINE1
$70 \leq x < 75$	6	OsSN2.1, p-SINE2, PoaS-IX, PoaS-X.3, PoaS-XI.3, ZmSINE2.1
$x < 70$	2	PoaS-V.1, PoaS-XIII

[a] of SINE families

Table S6. Average length of target site duplications and 3' tail of Poaceae SINE families.

SINE family	Species	Copy number[a]	TSD			3' tail		
			Copies[b]	Average[c]	Percent[d]	Copies[e]	Average[f]	Percent[g]
Au	*T. aestivum*	471	397	14	15.7	250	7	46.9
OsSN1	*O. sativa*	89	73	12	18.0	77	9	13.5
OsSN2.1	*O. sativa*	97	74	11	23.7	83	9	14.4
OsSN2.2	*T. aestivum*	541	430	10	20.5	438	7	19.0
OsSN3	*O. sativa*	168	126	9	25.0	142	9	15.5
p-SINE1	*O. sativa*	631	558	12	11.6	470	8	25.5
p-SINE2	*O. sativa*	31	28	12	9.7	20	8	35.5
p-SINE3	*O. sativa*	24	23	12	4.2	18	8	25.0
PoaS-I	*B. distachyon*	90	78	12	13.3	76	8	15.6
PoaS-II	*B. distachyon*	126	111	11	11.9	111	8	11.9
PoaS-III	*P. virgatum*	27	26	12	3.7	26	9	3.7
PoaS-IV	*O. sativa*	10	9	11	10.0	8	8	20.0
PoaS-V.1	*S. bicolor*	62	58	9	6.5	43	8	30.6
PoaS-V.2	*H. vulgare*	223	155	11	30.5	199	7	10.8
PoaS-VI	*T. aestivum*	376	351	12	6.6	359	8	4.5
PoaS-VII	*O. sativa*	11	11	12	0.0	10	9	9.1
PoaS-VIII	*P. virgatum*	9	8	15	11.1	9	10	0.0
PoaS-IX	*O. sativa*	210	185	10	11.9	164	8	21.9
PoaS-X.1	*T. aestivum*	53	47	16	11.3	47	9	11.3
PoaS-X.2	*T. aestivum*	45	42	13	6.7	22	8	51.1
PoaS-X.3	*T. aestivum*	59	52	10	11.9	41	7	30.5
PoaS-XI.1	*T. aestivum*	244	229	12	6.1	204	8	16.4
PoaS-XI.2	*T. aestivum*	67	59	11	11.9	53	7	20.9
PoaS-XI.3	*T. aestivum*	24	24	12	0.0	22	9	8.3
PoaS-XII	*S. bicolor*	47	38	12	19.1	41	8	12.8
PoaS-XIII	*O. sativa*	129	106	9	17.8	84	9	34.9
PoaS-XIV	*T. aestivum*	1	1	9	0.0	1	9	0.0
ZmSINE1	*Z. mays*	294	270	10	8.2	134	7	54.4
ZmSINE2.1	*P. virgatum*	52	48	14	7.7	44	8	15.4
ZmSINE2.2	*Z. mays*	11	6	6	45.5	9	7	18.2
ZmSINE2.3	*Z. mays*	34	20	8	41.2	27	7	20.6
ZmSINE3	*S. bicolor*	41	38	14	7.3	33	9	19.5
total[h]		**4,297**	**3,681**		**14.3**	**3,265**		**24.0**

[a] full-length copies
[b] with a detectable TSD
[c] average TSD length of the SINE family
[d] percentage of full-length copies without detectable TSD
[e] with a detectable 3' tail
[f] average tail length of the SINE family
[g] percentage of full-length copies without detectable 3' tail
[h] copy number and percentage of full-length copies without detectable TSD and 3' tail, respectively

Table S7. Analyzed plant SINE families with regard to the position of A and B box motif.

#	SINE family	Reference	#	SINE family	Reference	#	SINE family	Reference
1	AmaS-I	Schwichtenberg *et al.*, 2016	36	FabaS-VIII (LJ_SINE-1)	Gadzalski and Sakowicz, 2011	71	SaliS-V	Wenke *et al.*, 2011
2	AmaS-IIa	Schwichtenberg *et al.*, 2016	37	LJ_SINE-2	Gadzalski and Sakowicz, 2011	72	SB1 (S1Bn)	Deragon and Zhang, 2006
3	AmaS-III	Schwichtenberg *et al.*, 2016	38	LJ_SINE-3	Gadzalski and Sakowicz, 2011	73	SB2 (RathE1)	Deragon and Zhang, 2006
4	AmaS-IVa	Schwichtenberg *et al.*, 2016	39	NymS-I	Wenke *et al.*, 2011	74	SB3 (RathE2)	Deragon and Zhang, 2006
5	AmaS-IX	Schwichtenberg *et al.*, 2016	40	OsSN1	Tsuchimoto *et al.*, 2008	75	SB4 (RathE3)	Deragon and Zhang, 2006
6	AmaS-V	Schwichtenberg *et al.*, 2016	41	OsSN2.1	Tsuchimoto *et al.*, 2008	76	SB5	Deragon and Zhang, 2006
7	AmaS-VIa	Schwichtenberg *et al.*, 2016	42	OsSN2.2	this study	77	SB6	Deragon and Zhang, 2006
8	AmaS-VII	Schwichtenberg *et al.*, 2016	43	OsSN3	Tsuchimoto *et al.*, 2008	78	SB7	Deragon and Zhang, 2006
9	AmaS-VIII	Schwichtenberg *et al.*, 2016	44	PinS-I	Wenke *et al.*, 2011	79	SB8	Deragon and Zhang, 2006
10	AmaS-X	Schwichtenberg *et al.*, 2016	45	PoaS-I	Wenke *et al.*, 2011	80	SB9	Deragon and Zhang, 2006
11	AmaS-XI	Schwichtenberg *et al.*, 2016	46	PoaS-II	Wenke *et al.*, 2011	81	SB10	Deragon and Zhang, 2006
12	AmaS-XII	Schwichtenberg *et al.*, 2016	47	PoaS-III	this study	82	SB11	Deragon and Zhang, 2006
13	AmaS-XIII	Schwichtenberg *et al.*, 2016	48	PoaS-IV	this study	83	SB12	Deragon and Zhang, 2006
14	AmaS-XIV	Schwichtenberg *et al.*, 2016	49	PoaS-IX	this study	84	SB13	Deragon and Zhang, 2006
15	AmaS-XIX	Schwichtenberg *et al.*, 2016	50	PoaS-V.1	this study	85	SB14	Deragon and Zhang, 2006
16	AmaS-XV	Schwichtenberg *et al.*, 2016	51	PoaS-V.2	this study	86	SB15	Deragon and Zhang, 2006
17	AmaS-XVI	Schwichtenberg *et al.*, 2016	52	PoaS-VI	this study	87	ScroS-I	Wenke *et al.*, 2011
18	AmaS-XVII	Schwichtenberg *et al.*, 2016	53	PoaS-VII	this study	88	SolS-I	Wenke *et al.*, 2011
19	AmaS-XVIII	Schwichtenberg *et al.*, 2016	54	PoaS-VIII	this study	89	SolS-II	Wenke *et al.*, 2011
20	AmaS-XX	Schwichtenberg *et al.*, 2016	55	PoaSX.1	this study	90	SolS-III	Wenke *et al.*, 2011
21	AmaS-XXI	Schwichtenberg *et al.*, 2016	56	PoaSX.2	this study	91	SolS-IV	Wenke *et al.*, 2011
22	Au	Yagi *et al.*, 2011	57	PoaSX.3	this study	92	SolS-IX	Wenke *et al.*, 2011
23	BraS-I	Wenke *et al.*, 2011	58	PoaSXI.1	this study	93	SolS-V	Wenke *et al.*, 2011
24	CucuS-I	Wenke *et al.*, 2011	59	PoaSXI.2	this study	94	SolS-VI	Wenke *et al.*, 2011
25	CucuS-II	Wenke *et al.*, 2011	60	PoaSXI.3	this study	95	SolS-VII	Wenke *et al.*, 2011
26	CypS-I	Wenke *et al.*, 2011	61	PoaSXII	this study	96	SolS-VIII	Wenke *et al.*, 2011
27	EuphS-I	Wenke *et al.*, 2011	62	PoaSXIII	this study	97	TS	Yoshioka *et al.*, 1993
28	FabaS-I (MT_SINE-1)	Gadzalski and Sakowicz, 2011	63	PoaS-XIV	this study	98	VitaS-I	Wenke *et al.*, 2011
29	FabaS-II	Wenke *et al.*, 2011	64	p-SINE1	Mochizuki *et al.*, 1992	99	ZmSINE1	Baucom *et al.*, 2009
30	FabaS-III (MT_SINE-3)	Gadzalski and Sakowicz, 2011	65	p-SINE2	Xu *et al.*, 2005	100	ZmSINE2.1	Baucom *et al.*, 2009
31	FabaS-IV	Wenke *et al.*, 2011	66	p-SINE3	Xu *et al.*, 2005	101	ZmSINE2.2	Baucom *et al.*, 2009
32	FabaS-IX	Wenke *et al.*, 2011	67	SaliS-I	Wenke *et al.*, 2011	102	ZmSINE2.3	Baucom *et al.*, 2009
33	FabaS-V	Wenke *et al.*, 2011	68	SaliS-II	Wenke *et al.*, 2011	103	ZmSINE3	Baucom *et al.*, 2009
34	FabaS-VI	Wenke *et al.*, 2011	69	SaliS-III	Wenke *et al.*, 2011			
35	FabaS-VII (MT_SINE-2)	Gadzalski and Sakowicz, 2011	70	SaliS-IV	Wenke *et al.*, 2011			

Table S8. Transcribed SINE families of the wheat genome. The number of SINE transcripts for each wheat SINE family was obtained by NCBI megablast searches in the transcriptome shotgun assembly of *Triticum aestivum* using the respective PoaS consensus sequences as queries. Only full-length or near full-length hits (query coverage of at least 80 %) are included.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
Au	JP826147.1	99.5	98.3
(241 hits)	JV887736.1	98.9	99.4
	JP824021.1	99.4	96.6
	JP855580.1	99.4	96.1
	GFFI01000850.1	99.4	96.6
	GBKK01001488.1	98.4	98.9
	JP826530.1	98.4	98.3
	JP823777.1	97.9	100.0
	JV866695.1	97.9	100.0
	GFFI01001370.1	97.9	100.0
	GAJL01152802.1	97.9	100.0
	GAJL01055448.1	97.9	100.0
	JP826012.1	97.3	100.0
	JP824657.1	98.9	96.1
	GFFI01053330.1	99.4	94.9
	GFFI01033169.1	98.9	96.6
	GDTJ01001313.1	97.8	99.4
	GBKH01000584.1	98.4	97.8
	JP826059.1	97.8	98.3
	JP887289.1	99.4	93.8
	JP826270.1	99.4	93.8
	JP825990.1	97.3	99.4
	JP825278.1	97.3	100.0
	JP824794.1	97.3	100.0
	JV864562.1	97.3	100.0
	HP620901.1	97.8	98.3
	HP620900.1	97.8	98.3
	HP620899.1	97.8	98.3
	GFFI01001770.1	99.4	93.8
	JP823939.1	97.3	99.4
	GFFI01001999.1	97.3	99.4
	GFFI01284513.1	86.6	86.0
	GFFI01071017.1	85.1	96.6
	GFFI01165901.1	85.5	99.4
	GFFI01041102.1	89.5	80.3
	GFFI01083910.1	86.5	98.9
	GFFI01020351.1	90.1	80.3

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	GFFI01063784.1	85.9	97.2
	GFFI01006899.1	87.0	98.3
	GFFI01015155.1	88.5	92.7
	GFFI01038122.1	83.0	96.6
	GFFI01012793.1	88.4	96.6
	GFFI01001196.1	88.1	94.4
	GFFI01141314.1	89.1	97.8
	GFFI01195332.1	91.3	85.4
	GAEF01114790.1	93.5	81.5
	GFFI01018427.1	93.0	83.1
	GFFI01005468.1	89.1	98.3
	GFFI01208231.1	90.1	96.6
	GFFI01067803.1	90.2	96.6
	GFFI01077742.1	90.1	96.6
	GFFI01130466.1	90.1	96.6
	GFFI01049210.1	90.3	98.9
	GFFI01009649.1	93.5	89.9
	GFFI01012261.1	90.6	98.3
	GFFI01161085.1	94.9	83.7
	GFFI01190510.1	91.7	96.6
	GFFI01060856.1	91.7	96.6
	GFFI01156263.1	91.7	96.6
	HAAB01049874.1	91.7	96.6
	GFFI01100515.1	91.4	98.9
	GFFI01192564.1	91.8	97.8
	GFFI01016502.1	95.1	87.1
	GFFI01032016.1	96.8	83.1
	GFFI01051290.1	96.3	85.4
	GFFI01018704.1	96.3	86.0
	GFFI01097188.1	96.3	86.0
	GFFI01129490.1	92.5	99.4
	GFFI01184775.1	95.3	89.3
	GFFI01038300.1	95.3	89.9
	GFFI01001564.1	96.9	86.5
	GFFI01046159.1	93.9	96.6
	GFFI01118585.1	94.4	96.1
	GFFI01018210.1	95.4	93.3
	GFFI01112786.1	94.5	96.1
	GFFI01197180.1	95.0	94.9
	GFFI01000613.1	94.5	97.2
	GFFI01001410.1	94.5	97.2

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	GFFI01160410.1	94.6	98.3
	GFFI01012844.1	95.0	96.6
	GFFI01073908.1	96.5	91.6
	GFFI01001809.1	97.6	89.9
	GFFI01037936.1	97.6	89.9
	GFFI01012100.1	95.2	99.4
	GFFI01051109.1	95.6	97.8
	GFFI01010060.1	97.1	93.3
	GFFI01011377.1	96.1	96.6
	GFFI01026090.1	96.6	94.9
	GFFI01033655.1	96.6	95.5
	GFFI01141220.1	95.7	98.3
	HAAB01000019.1	98.8	88.2
	HAAB01040679.1	96.6	94.9
	HAAB01040680.1	96.6	94.9
	HAAB01040681.1	96.6	94.9
	GFFI01005162.1	97.2	93.8
	GFFI01003639.1	96.2	97.8
	GFFI01006222.1	95.7	100.0
	GFFI01040058.1	96.2	97.8
	GFFI01003030.1	96.2	98.3
	GFFI01038694.1	97.2	94.9
	GFFI01075387.1	98.8	89.9
	HAAB01018502.1	97.2	94.9
	HAAB01018504.1	97.2	94.9
	HAAB01046637.1	97.2	94.9
	GFFI01000987.1	96.7	97.8
	GFFI01050879.1	97.2	96.1
	GFFI01001643.1	98.3	93.3
	GFFI01004121.1	97.2	96.6
	GFFI01018945.1	96.7	98.3
	GFFI01022373.1	97.2	96.6
	GFFI01005509.1	97.3	97.2
	GBKI01003371.1	99.4	91.6
	GFFI01004173.1	97.8	96.6
	GFFI01023144.1	97.3	97.8
	GFFI01185971.1	96.8	99.4
	GBZP01015204.1	97.8	97.2
	GFFI01005804.1	98.3	95.5
	GFFI01011581.1	96.8	100.0
	GFFI01019511.1	96.8	100.0

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	GFFI01023884.1	96.8	100.0
	GFFI01005349.1	97.8	97.2
	GFFI01007775.1	97.3	98.9
	GFFI01008219.1	97.3	99.4
	GFFI01015053.1	98.9	93.8
	GBKI01001698.1	97.8	97.8
	JW033136.1	97.3	99.4
	JP826171.1	97.3	97.8
	JP825772.1	98.3	94.9
	JV951105.1	97.8	96.6
	HP620898.1	97.3	98.3
	HP620897.1	97.3	98.3
	HP620896.1	97.3	98.3
	JP208898.1	97.3	98.3
	JP906232.1	97.3	97.8
	JP825993.1	97.3	97.8
	JP825128.1	96.8	99.4
	JV865544.1	99.4	91.6
	JP208929.1	97.3	97.8
	JP845343.1	96.8	98.9
	JP825700.1	97.2	96.1
	JP825220.1	97.3	97.2
	JP824076.1	96.3	100.0
	JP866573.1	96.2	99.4
	JP823981.1	97.3	96.6
	JP823964.1	97.2	96.6
	JW031317.1	97.3	96.6
	JP850485.1	95.8	100.0
	JP826472.1	95.8	100.0
	JP826396.1	96.7	97.8
	JP825119.1	96.7	96.6
	JV871277.1	96.7	97.8
	JV863616.1	96.7	97.8
	JP882381.1	95.8	100.0
	JP825858.1	97.2	95.5
	JP825596.1	97.2	94.9
	JP820507.1	98.3	92.1
	JV865078.1	96.7	97.2
	HAAB01018502.1	97.2	94.9
	HAAB01018504.1	97.2	94.9
	HAAB01046637.1	97.2	94.9

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	JP941373.1	98.3	91.6
	JP826177.1	97.2	94.9
	JP826158.1	96.2	98.3
	JP824473.1	96.2	98.3
	JP824083.1	98.8	89.9
	JP823819.1	96.2	98.3
	JV890302.1	98.8	89.9
	JV887195.1	98.8	89.9
	JP919055.1	95.7	99.4
	JP855304.1	97.2	93.3
	JP847838.1	96.7	96.6
	JP826550.1	95.7	98.3
	JP826333.1	97.2	93.3
	JP826239.1	96.1	96.6
	JP826067.1	95.7	99.4
	JP826061.1	96.6	94.9
	JP825182.1	95.7	97.8
	JP826297.1	96.2	96.6
	JP826011.1	98.2	89.9
	HAAB01000019.1	98.8	88.2
	HAAB01040679.1	96.6	94.9
	HAAB01040680.1	96.6	94.9
	HAAB01040681.1	96.6	94.9
	JP893661.1	96.6	94.9
	JP826422.1	97.1	93.3
	JP826135.1	97.7	91.6
	JP825502.1	98.2	89.9
	JV911066.1	96.6	94.9
	HP627074.1	98.2	89.9
	JP208899.1	97.1	93.3
	JP826424.1	97.1	92.1
	JP826336.1	95.1	96.1
	JP825180.1	95.6	96.6
	JP826376.1	96.1	94.9
	JV924743.1	95.6	97.2
	JP826331.1	95.1	98.3
	JP824489.1	94.5	96.6
	JP826269.1	96.0	93.8
	JP826190.1	96.1	93.8
	JV835478.1	94.6	99.4
	JP825462.1	96.5	91.6

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	JP825249.1	93.6	99.4
	JV907689.1	95.1	97.2
	JP825565.1	97.6	87.6
	JP826218.1	95.5	92.7
	JV883811.1	97.6	87.6
	JP855302.1	97.0	88.2
	JP826292.1	94.8	92.7
	JV869258.1	94.5	97.2
	JP826245.1	96.4	87.6
	JV940911.1	97.6	86.5
	JV905219.1	94.1	97.8
	JP915271.1	96.4	89.3
	JP826303.1	97.5	86.0
	JW030607.1	96.4	89.3
	JP826466.1	93.2	100.0
	JV915936.1	94.9	94.4
	JP826414.1	95.7	86.0
	JP824582.1	96.9	86.5
	JV866433.1	96.9	86.5
	JP824085.1	96.9	84.3
	JP824245.1	93.6	91.0
	JV853274.1	91.4	98.9
	HAAB01049874.1	91.7	96.6
	JV844441.1	96.1	80.9
	JV899083.1	90.6	98.3
	JP824893.1	94.9	83.1
	JP902249.1	90.3	98.9
	JP826548.1	90.3	98.9
	JV925406.1	90.3	98.9
	JV894974.1	90.3	98.9
	JV862087.1	89.1	98.3
	JV897834.1	93.0	83.1
	JV898470.1	93.5	80.9
	JP826320.1	88.8	94.9
	JP826151.1	90.3	91.6
	JW003572.1	89.7	92.1
	JV940443.1	92.8	80.9
	JV815865.1	91.3	85.4
	JP220237.1	89.7	91.6
	JW008386.1	88.1	94.4

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
OsSN2.2	GDTJ01001833.1	92.2	100.0
(47 hits)	JW026700.1	92.1	100.0
	GAJL01262256.1	92.1	100.0
	GAJL01260279.1	92.1	100.0
	GBKI01001427.1	93.1	95.1
	GAJL01278844.1	91.8	100.0
	GAJL01278733.1	91.8	100.0
	GAJL01278363.1	91.8	100.0
	GAJL01278272.1	91.8	100.0
	GAEF01013474.1	91.2	100.0
	JV865090.1	90.7	99.3
	GFFI01102921.1	90.2	100.0
	JP914768.1	90.1	98.9
	JP906781.1	89.9	100.0
	HP619629.1	90.0	98.9
	GFFI01009683.1	90.0	98.6
	GBZP01000908.1	89.7	99.6
	HP619628.1	90.0	95.4
	JV895444.1	89.1	99.3
	GFFI01171427.1	89.1	99.3
	JV920434.1	88.7	99.3
	HP622491.1	88.5	100.0
	HP619627.1	88.7	98.9
	JV989566.1	88.4	99.3
	GFFI01108368.1	88.4	99.3
	JP906782.1	88.2	98.9
	GFFI01204204.1	90.4	88.7
	JV990452.1	87.9	100.0
	JP906778.1	87.8	94.3
	HP619626.1	88.7	95.4
	JP207269.1	87.9	99.6
	HP617884.1	88.7	92.9
	HP617882.1	88.7	92.9
	HP617880.1	88.7	92.9
	HP619631.1	87.5	97.9
	GFFI01149605.1	86.8	98.9
	HP619630.1	87.5	94.3
	GAJL01227330.1	88.0	90.8
	GAJL01225582.1	88.0	90.8
	GAJL01221511.1	88.0	90.8
	GFFI01113274.1	86.2	98.9

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	JV925751.1	85.5	94.7
	GFFI01226731.1	86.2	82.7
	GBZP01003092.1	84.1	96.8
	JV936651.1	85.7	83.0
	GAJL01208935.1	81.1	96.1
	GFFI01208503.1	80.5	88.3
p-SINE3		No significant similarity found	
PoaS-II	GAJL01073430.1	85.3	82.0
(2 hits)	GAEF01125074.1	85.3	82.0
PoaS-V.2	GFFI01114760.1	89.6	89.8
(2 hits)	GAJL01136355.1	88.7	89.8
PoaS-VI	JV886974.1	98.6	100.0
(29 hits)	GFFI01028534.1	98.6	100.0
	GAEF01025463.1	98.6	100.0
	HAAB01023782.1	97.2	100.0
	HAAB01023783.1	97.2	100.0
	JP225370.1	97.9	97.8
	HAAB01023783.1	97.2	100.0
	HAAB01023782.1	97.2	100.0
	GFFI01047454.1	97.2	100.0
	HP624347.1	96.5	100.0
	GFFI01001601.1	95.2	100.0
	GAJL01079605.1	95.1	100.0
	JP826320.1	95.1	98.5
	JW019124.1	97.7	90.3
	GFFI01106819.1	97.7	90.3
	JV953615.1	96.9	90.3
	JV828468.1	96.9	90.3
	JV966810.1	93.7	99.3
	GFFI01226037.1	93.7	99.3
	GFFI01123385.1	95.4	90.3
	GFFI01229186.1	94.7	91.8
	HAAB01023785.1	91.4	97.0
	JP900339.1	91.4	97.0
	JV911546.1	91.4	97.0
	HAAB01023785.1	91.4	97.0
	GFFI01070040.1	91.4	97.0
	GDTJ01006264.1	91.4	97.0
	GFFI01124338.1	88.7	99.3
	GFFI01254088.1	87.1	91.8

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
PoaS-X.1	GAEF01035603.1	100.0	100.0
(11 hits)	HAAB01013044.1	99.4	100.0
	HAAB01013044.1	99.4	100.0
	GFFI01048184.1	99.4	100.0
	JV846483.1	98.8	100.0
	GFFI01127781.1	98.8	100.0
	GAJL01106061.1	98.8	100.0
	JP934915.1	99.4	96.1
	GAJL01024900.1	98.2	100.0
	JP934916.1	98.7	97.4
	GAEF01086796.1	100.0	83.1
PoaS-X.2	GFFI01003491.1	98.8	100.0
(11 hits)	GAEF01003102.1	100.0	94.7
	HAAB01076462.1	99.3	94.7
	HAAB01076462.1	99.3	94.7
	GFFI01011092.1	99.4	94.7
	JW031214.1	98.7	95.4
	GFFI01176110.1	98.7	95.4
	GFFI01004271.1	98.1	95.4
	JP223316.1	97.8	84.2
	JW018229.1	94.1	82.9
	GFFI01256141.1	94.1	82.9
PoaS-X.3	GFFI01003491.1	78.9	100.0
(7 hits)	GAEF01003102.1	80.4	88.7
	HAAB01076462.1	79.9	89.3
	HAAB01076462.1	79.9	89.3
	GFFI01011092.1	79.9	89.3
	JW031214.1	79.2	89.3
	GFFI01004271.1	79.2	89.3
PoaS-XI.1	GFFI01097706.1	95.4	100.0
(21 hits)	JW011883.1	95.4	99.3
	GBZP01029029.1	94.7	98.6
	GFFI01003751.1	95.2	95.8
	GFFI01114353.1	94.1	98.6
	GAJL01262623.1	94.0	97.9
	JV964709.1	92.8	99.3
	GFFI01213080.1	92.8	99.3
	GFFI01118501.1	92.8	99.3
	GFFI01045420.1	95.6	88.9
	GFFI01029491.1	95.6	88.9
	GFFI01017887.1	95.6	88.9

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	GFFI01014927.1	95.6	88.9
	GFFI01254360.1	92.8	99.3
	GAJL01260003.1	95.6	88.2
	GAJL01243314.1	95.6	88.2
	GAJL01217484.1	95.6	88.2
	GFFI01054649.1	94.3	91.0
	GAJL01173572.1	93.1	93.1
	HP631059.1	93.5	89.6
	GFFI01146364.1	90.5	95.8
PoaS-XI.2 (3 hits)	JV990572.1	90.3	99.3
	GFFI01013733.1	90.3	99.3
	JV851135.1	90.2	97.9
PoaS-XI.3 (1 hit)	GAJL01260867.1	87.3	100.0
PoaS-XII (3 hits)	JV831699.1	91.0	95.3
	GFFI01281887.1	91.0	95.3
	GAEF01108454.1	91.0	95.3
PoaS-XIV		No significant similarity found	
ZmSINE1 (93 hits)	JP845635.1	100.0	98.8
	JP845644.1	100.0	97.6
	JP845641.1	100.0	97.6
	JP825249.1	100.0	97.6
	HP633021.1	100.0	97.6
	GFFI01006748.1	100.0	97.6
	GAJL01262679.1	100.0	97.6
	HAAB01038319.1	99.4	98.8
	HAAB01038319.1	99.4	98.8
	GFFI01102213.1	100.0	97.0
	GFFI01016802.1	99.4	98.8
	GAJL01238346.1	100.0	97.0
	JW030112.1	100.0	96.4
	GFFI01013920.1	100.0	96.4
	GAEF01068183.1	100.0	96.4
	GAEF01068182.1	100.0	96.4
	GAEF01068180.1	99.4	98.8
	GAEF01028568.1	100.0	96.4
	GAEF01028567.1	100.0	96.4
	HAAB01080644.1	100.0	95.8
	JP845651.1	98.8	98.2
	JP845626.1	100.0	95.8

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	JV867328.1	100.0	95.8
	JP238566.1	100.0	95.8
	HAAB01080644.1	100.0	95.8
	GFFI01068670.1	100.0	95.8
	HAAB01025192.1	99.4	97.0
	HAAB01025192.1	99.4	97.0
	GFFI01011026.1	99.4	97.0
	JP879050.1	99.4	96.4
	JP845648.1	98.8	98.2
	JV906514.1	99.4	96.4
	JV849320.1	99.4	96.4
	GFFI01030804.1	99.4	96.4
	GBZP01002159.1	98.8	97.6
	HAAB01068557.1	99.4	95.8
	HAAB01068558.1	99.4	95.8
	JP845623.1	98.8	97.6
	JV890983.1	99.4	95.8
	HAAB01068558.1	99.4	95.8
	HAAB01068557.1	99.4	95.8
	GFFI01028724.1	99.4	95.8
	GDTJ01001488.1	98.8	97.6
	GBKH01002131.1	99.4	95.8
	JW030683.1	98.8	97.0
	JP845647.1	99.4	94.5
	HAAB01031833.1	98.2	97.6
	JP838077.1	98.8	95.8
	HAAB01031833.1	98.2	97.6
	GFFI01007224.1	98.2	97.6
	JV939536.1	98.8	95.2
	JV907364.1	98.8	95.2
	GFFI01098043.1	98.8	95.2
	JV866149.1	97.6	97.6
	GFFI01000690.1	98.8	94.5
	JP845649.1	98.2	94.5
	JP845636.1	97.0	95.8
	JW030843.1	99.4	90.3
	GFFI01068184.1	98.8	92.7
	JP845625.1	97.6	95.2
	HP635190.1	96.5	97.6
	GFFI01006886.1	99.4	88.5
	GFFI01162739.1	96.5	97.0

Table S8. Continued.

SINE family	Accession	Identity [%][a]	Coverage [%][b]
	GFFI01026362.1	100.0	86.1
	JP921844.1	98.7	88.5
	JP845643.1	98.7	88.5
	JP872385.1	96.4	94.5
	JV991813.1	98.7	87.9
	JP942965.1	100.0	80.6
	GFFI01095858.1	94.6	95.8
	GFFI01017981.1	92.4	98.2
	GFFI01102060.1	91.2	97.6
	JW029432.1	92.0	85.5
	GFFI01201640.1	92.0	85.5
	HAAB01001602.1	91.4	85.5
	HAAB01001602.1	91.4	85.5
	GFFI01097888.1	89.7	88.5
	GFFI01123104.1	84.7	97.6
	JV994851.1	85.5	95.2
	GFFI01091410.1	88.7	80.6
	GFFI01049822.1	84.9	95.2
	GFFI01020842.1	79.7	87.9
	GAJL01154297.1	80.8	81.8
	GAJL01245169.1	80.8	81.8
	GAJL01271261.1	80.8	81.8
	GBZP01001459.1	80.5	89.7
	HAAB01032894.1	80.5	89.7
	HAAB01032895.1	80.5	89.7
	HAAB01032896.1	80.5	89.7
	HAAB01032894.1	80.5	89.7
	HAAB01032895.1	80.5	89.7
	HAAB01032896.1	80.5	89.7
	JV821025.1	78.8	89.1

[a] to query

[b] without artificial tail sequence (9 bp)

Table S9. Potential promotor motifs of multimeric SINEs. Deviations from the conserved motif (box A – TAGCNCAG(N)TGG and box B - GGTTCGANNCC, Figure S5) are drawn in red color.

SINE family	1st unit (= 5' unit)		2nd unit		3rd unit	
	box A	box B	box A'	box B'	box A"	box B"
ZmSINE2.1 (*P. virgatum*)	TAGCCGGATTGG	GGTTCGACTCC	n.d.	GGTTCGGGGGT	n.d.	n.d.
ZmSINE2.2 (*Z. mays*)	TAGCTGAGTTGG	AGTTCGAATCC	TGGTTGTGTGC	GGCTCAAAGCA	TGTTGCACGGG	CGGTCGGGGCT
ZmSINE2.3 (*Z. mays*)	TAGCGTAATGG	GGTTCGATCCC	TTGCAGAGTGG	GGTTCGGGGAT	n.d.	n.d.
PoaS-XIII (*O. sativa*)	TGGTGTGGTGG	GGTTTAAATCC	TGTGCCGCTGG	CATTCGTGGGG	n.d.	n.d.
PoaS-XIV (*T. aestivum*)	TAGCCCAGTGG	AGTTCGATCCA	TAGCCCAGTGG	AGTTCGATCCA	n.d.	n.d.
PoaS-VII (*O. sativa*)	TTGCTCGGGTGG	CGTTCGAWCCC	GCGTTCAGTGG	TGTTAGGGACG	n.d.	n.d.
OsSN1 (*O. sativa*)	TAGCCTAGTGG	GGTTCGACTCC	n.d.	GGTTTGCGTGC	n.d.	n.d.
OsSN2.1 (*O. sativa*)	TAGCTCAACTGG	GGTTCAAATCC	n.d.	GGTGCTCATAG	n.d.	n.d.
OsSN2.2 (*T. aestivum*)	TGGTTGGATGG	GGTTCAAGTCC	n.d.	GGCTCAGTCTT	n.d.	n.d.

n.d. not detectable

Supplemental Information to

Chapter 2.3

Comparative analysis of SINEs in Salicaceae species reveals 3' end diversification in many families

Content

Supplemental Figures

Supplemental Tables

Supplemental Figures

SaliS-I

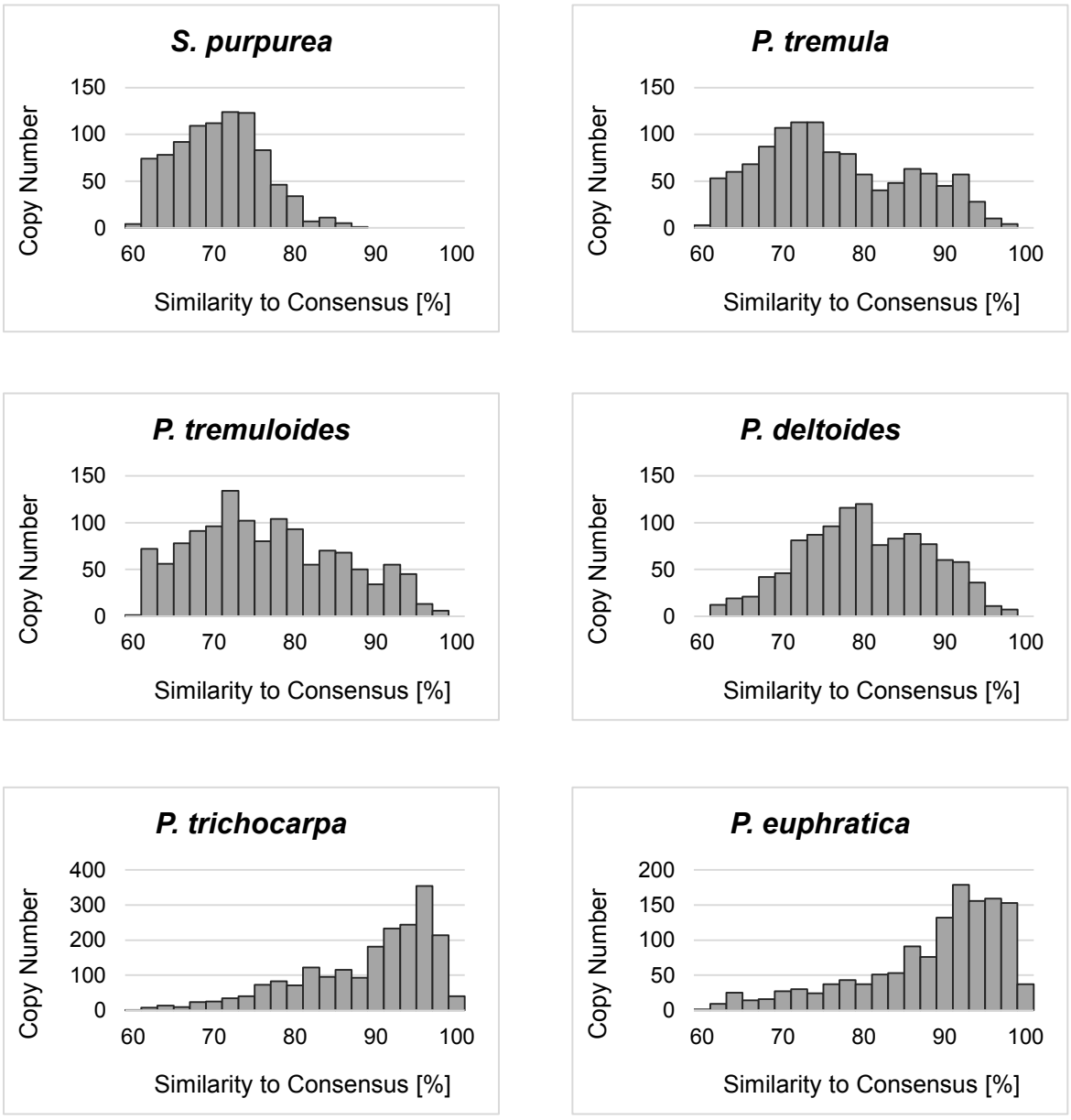
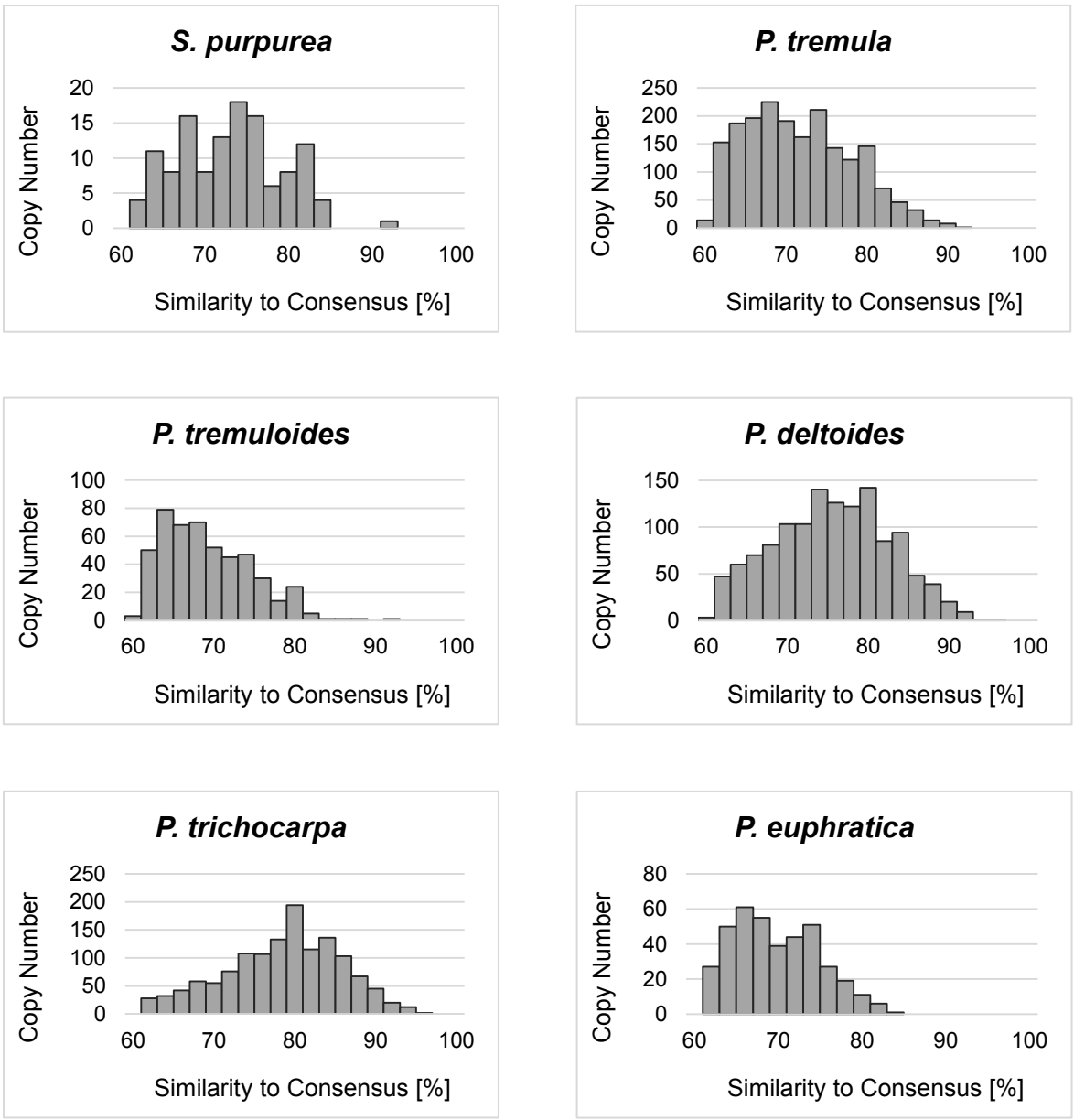


Figure S1. Activity profiles of SaliS families and subfamilies. All members of a SINE family were compared with the respective species-specific consensus sequence and the resulting percentage values were grouped into similarity intervals reflecting recent transpositional behavior and the relative age of copies. SaliS families and subfamilies are included, if the SINE family occurs with at least ten full-length copies.

SaliS-II



SaliS-III.1

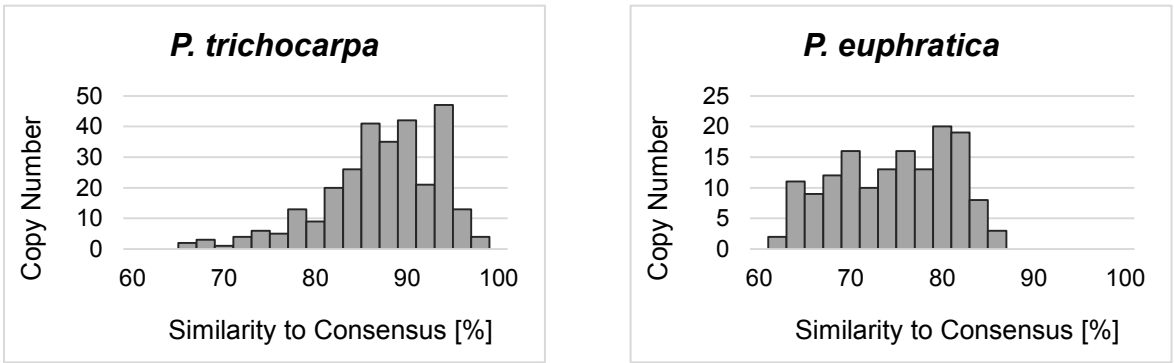
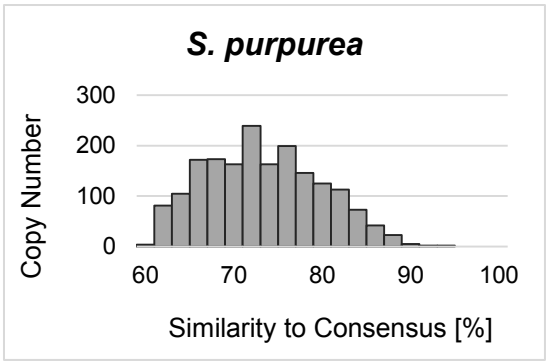
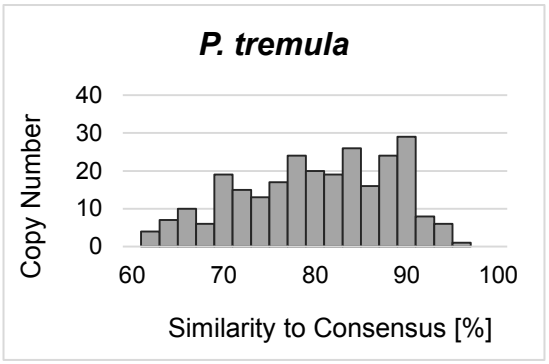
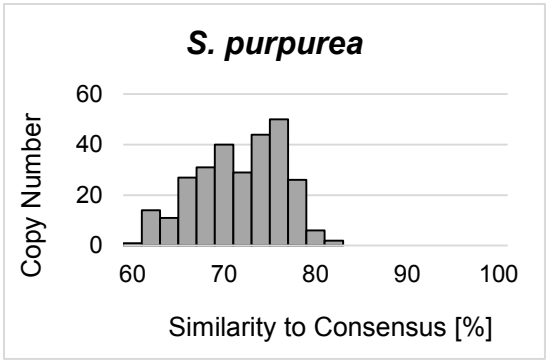


Figure S1. Continued.

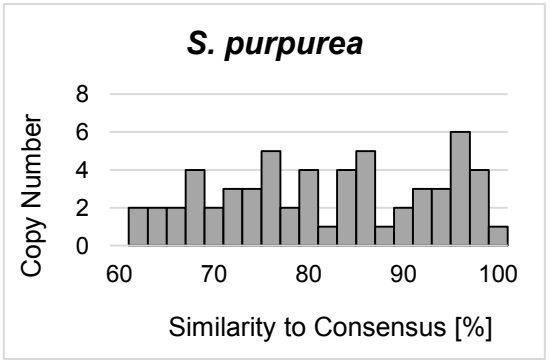




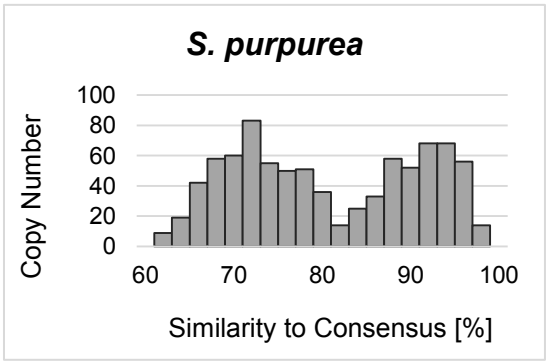
SaliS-III.2



SaliS-III.3



SaliS-IV.1



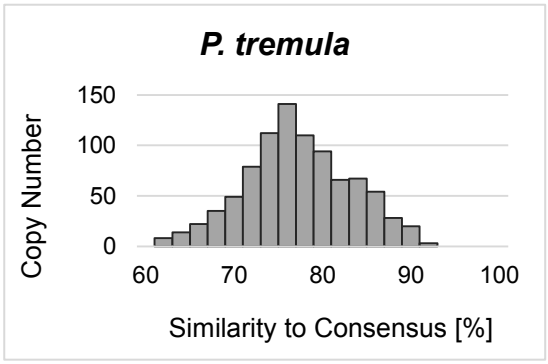
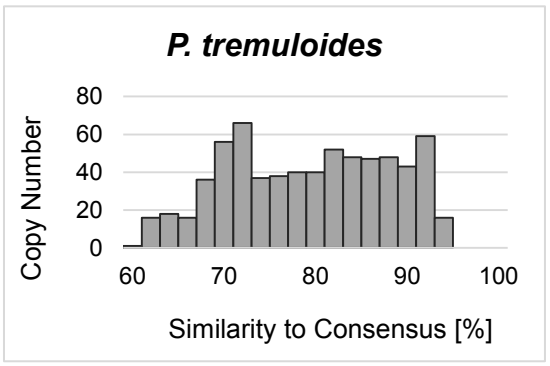
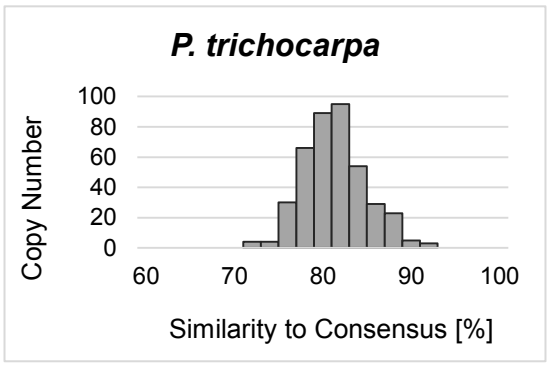
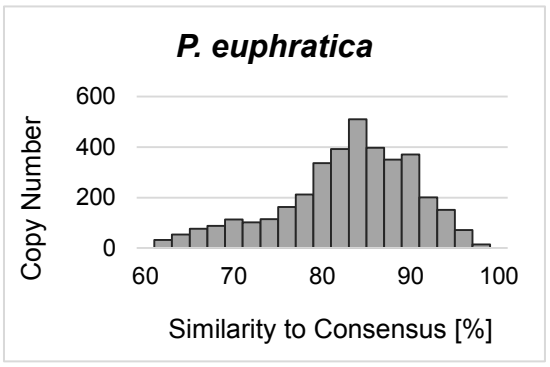


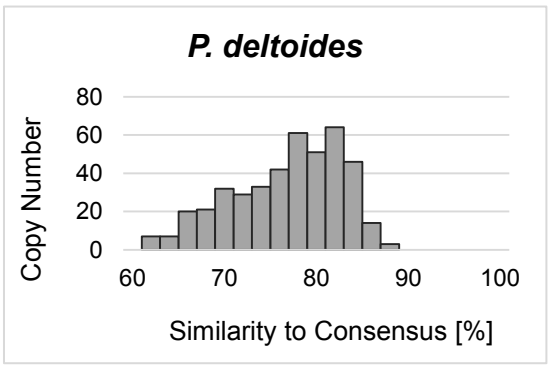
Figure S1. Continued.

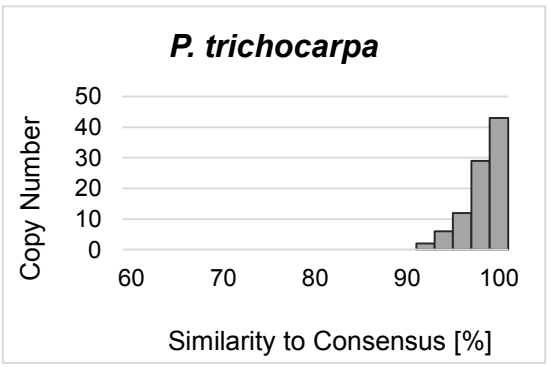




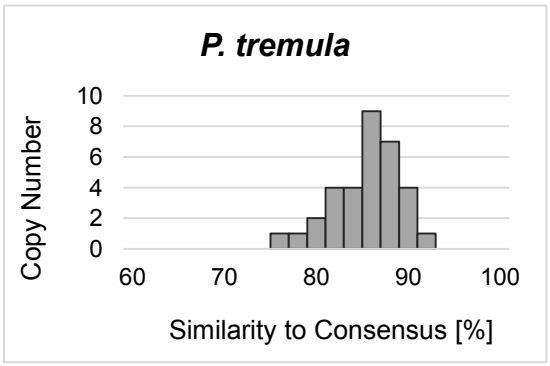


SaliS-IV.2





SaliS-IV.3



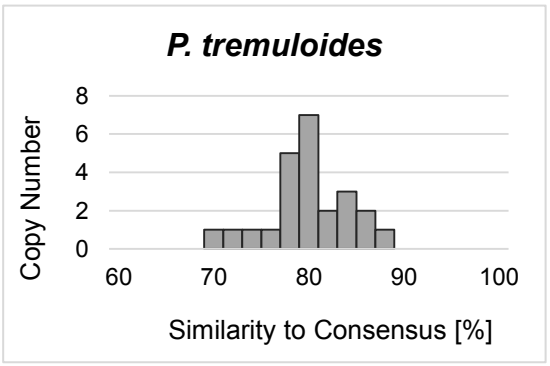
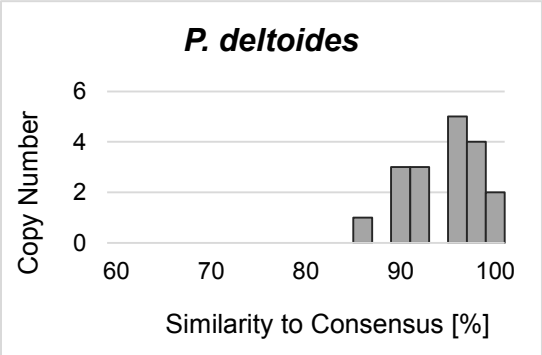
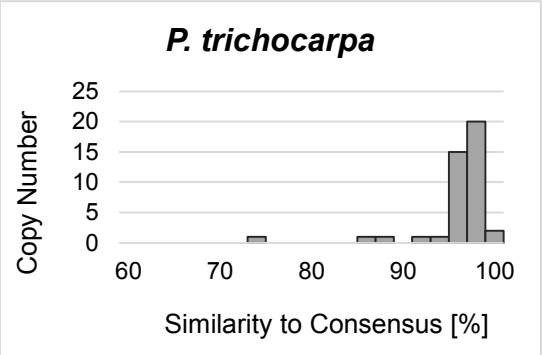


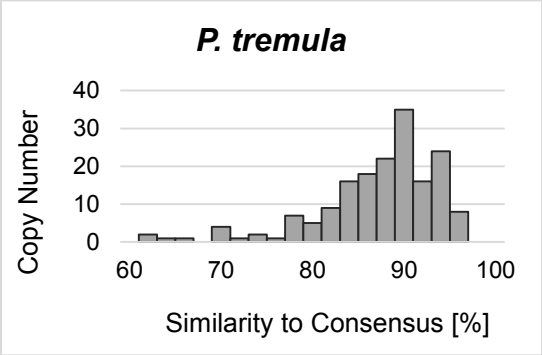
Figure S1. Continued.

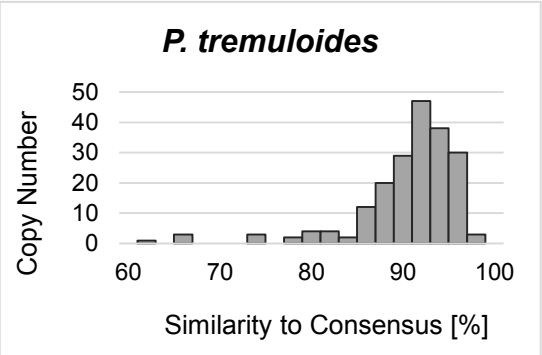
SaliS-V



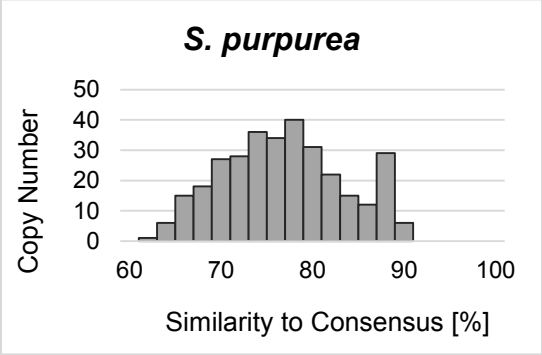


SaliS-VI.1





SaliS-VI.2



SaliS-VI.3

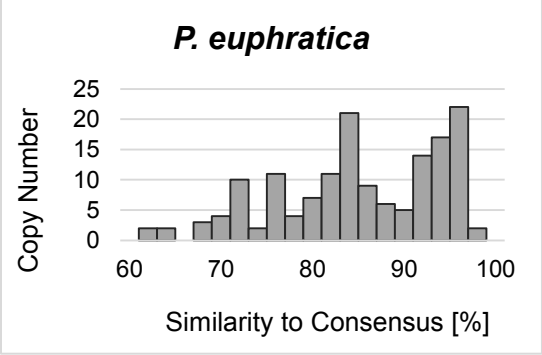
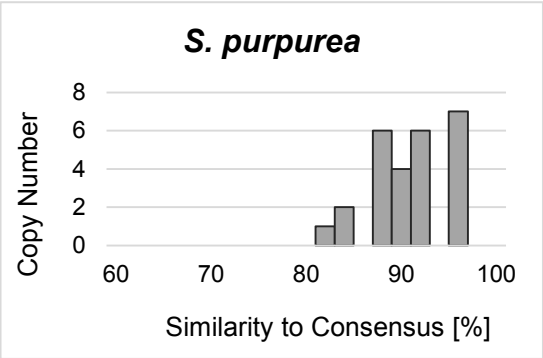
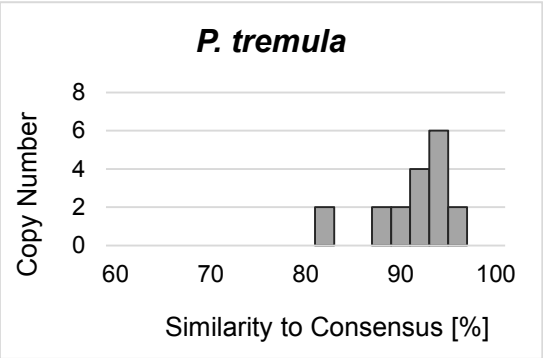
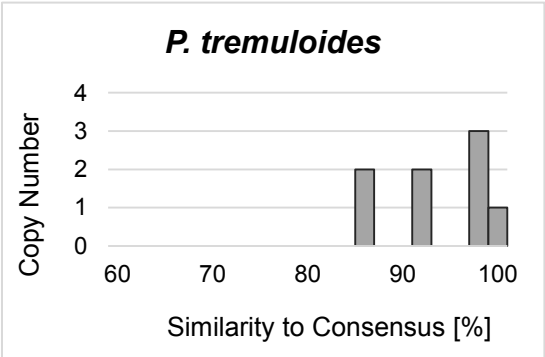


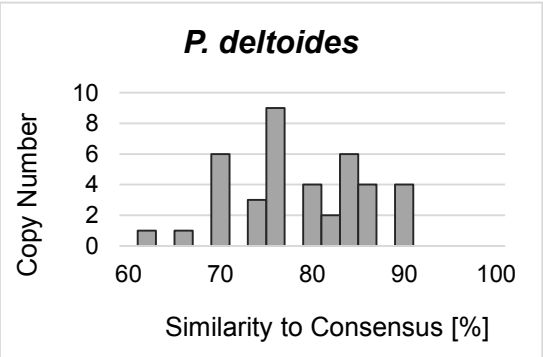
Figure S1. Continued.

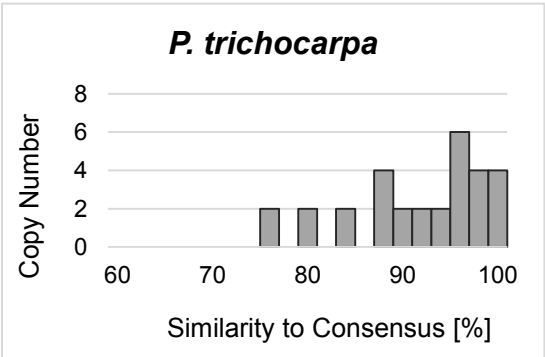
SaliS-VII.1



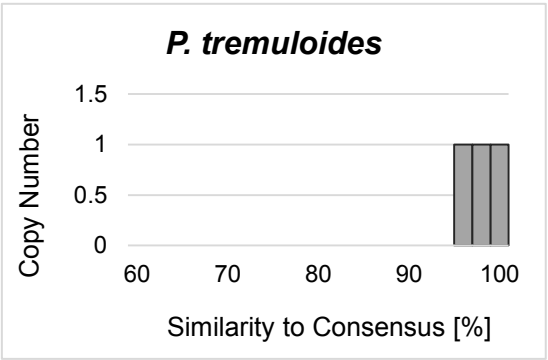








SaliS-VII.2



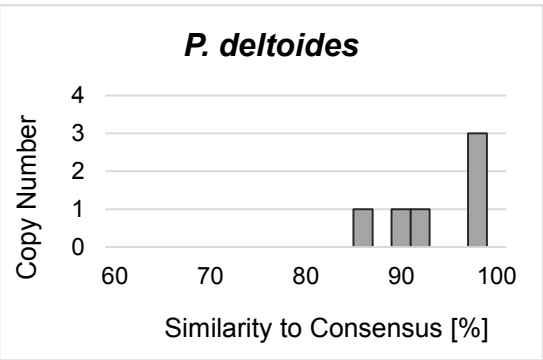
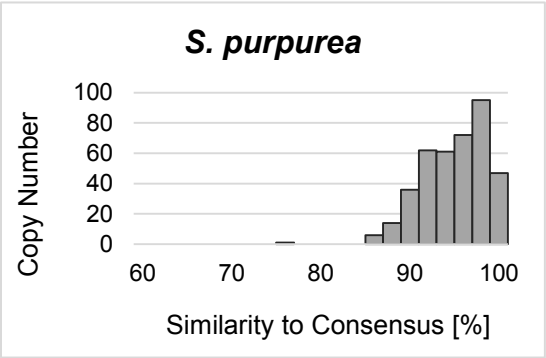
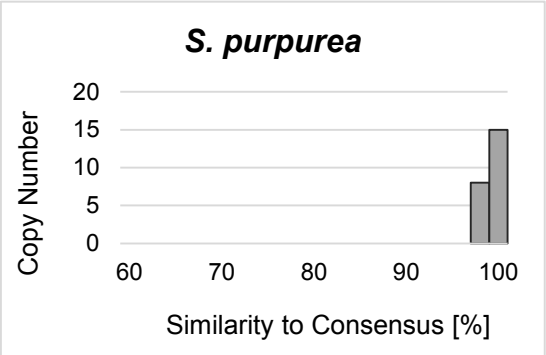


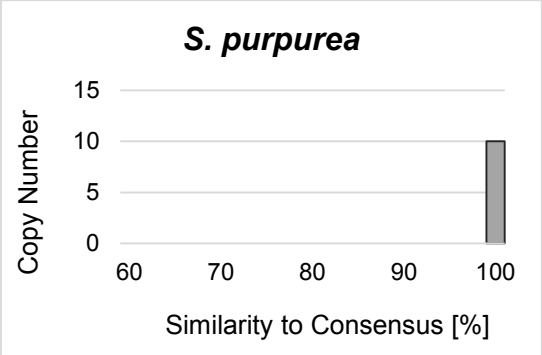
Figure S1. Continued.



SaliS-VII.3



SaliS-VII.4



SaliS-VIII

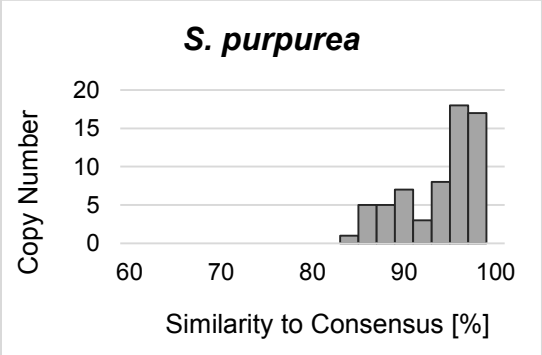
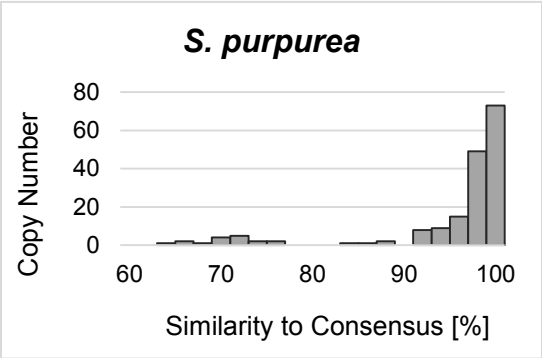
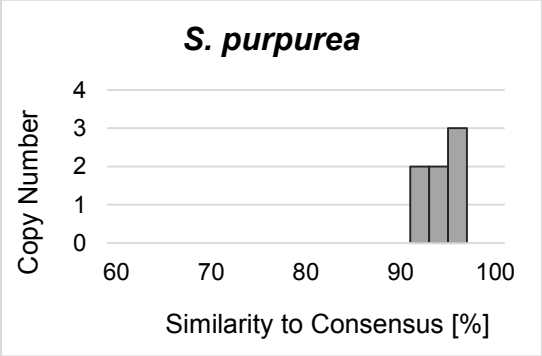


Figure S1. Continued.

SaliS-IX



SaliS-X



SaliS-XI

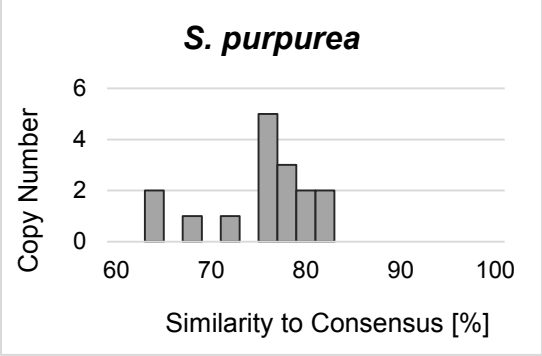


Figure S1. Continued.

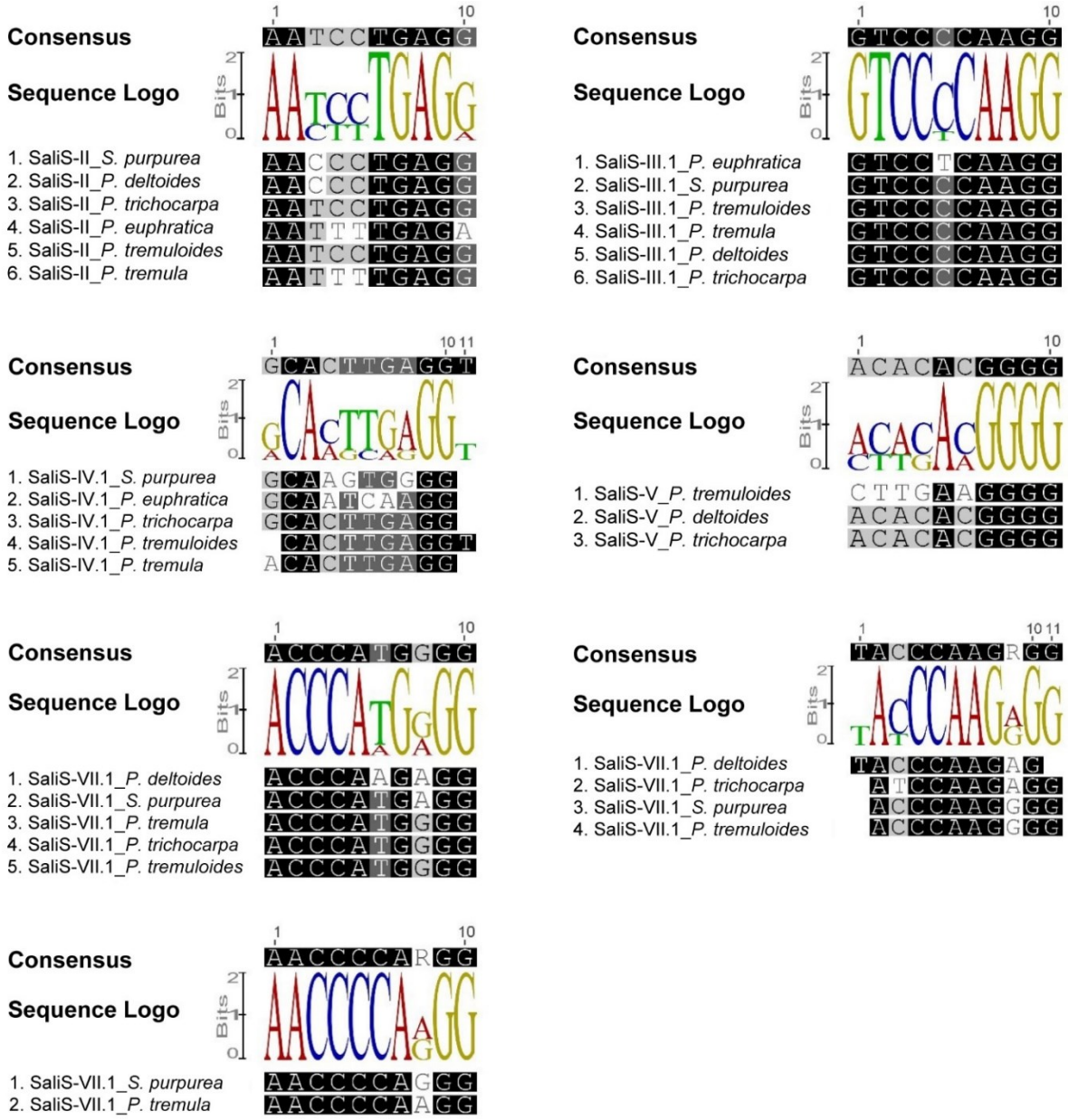


Figure S2. 5' start motifs of SaliS families and subfamilies of different species. Alignments and sequence logos comparing species-specific subfamily consensus sequences of the 5' start sequence are only shown in case of sequence polymorphism between species.

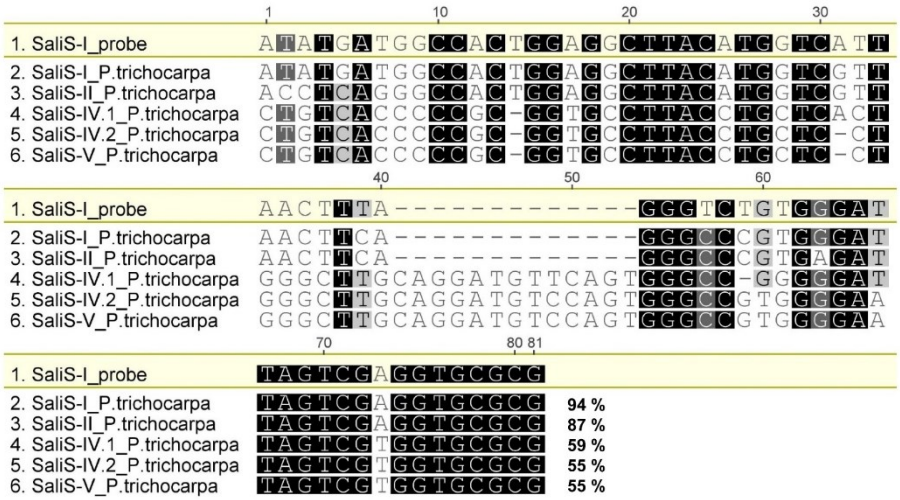


Figure S3. Comparison of the SaliS-I probe for fluorescent *in situ* hybridization with the respective region of the consensus sequences of SaliS families sharing the same 3' region (see Figure 2). The distances to the SaliS-I probe are provided at the 3' end.

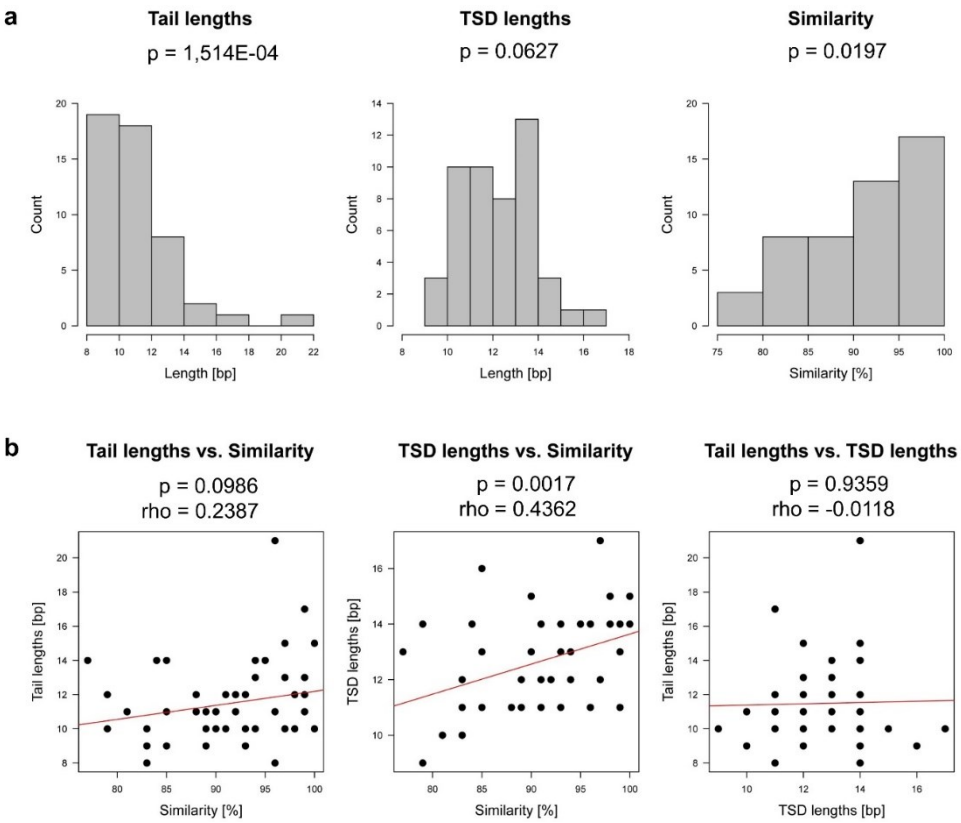


Figure S4. Correlation between TSD length, tail length and similarity. (a) Histograms showing the data distribution. Applying the Shapiro-Wilk test for normality, the similarity values are normally distributed, whereas tail lengths and TSD lengths are not (p-value <= 0.05). **(b)** Scatter plots illustrate the Spearman's rank correlation between the three SINE characteristics. The regression line is represented in red. TSD lengths and similarity values are significantly (p = 0.0017) positive correlated (rho = 0.4362), while no significant correlation was detected between tail lengths and TSD lengths (p = 0.9359) as well as between similarity values and tail lengths (p = 0.0986).

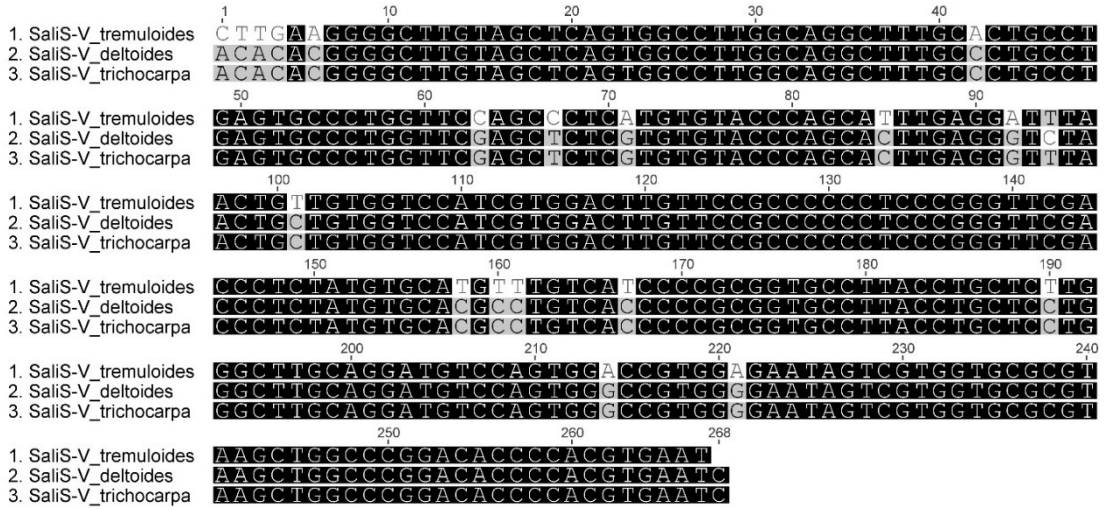


Figure S5. Comparison of SaliS-V consensus sequences in different poplar species. SaliS-V of *P. deltoides* and *P. trichocarpa* are highly similar (99.6 % consensus identity). The single SaliS-V copy found in *P. tremuloides* shares 92.5 % and 92.9 % similarity to the SaliS-V consensus sequences of *P. deltoides* and *P. trichocarpa*, respectively.

Supplemental Tables

Table S1. Sizes and sources of analyzed Salicaceae genomes.

Species	Source	Size [Mb]
Populus deltoides	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PdeltoidesWV94	446.7
Populus euphratica	https://www.ncbi.nlm.nih.gov/assembly/GCA_000495115.1	472.5
Populus tremula	ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.0/FASTA/	204.3
Populus tremuloides	ftp://plantgenie.org/Data/PopGenIE/Populus_tremuloides/v1.1/FASTA/	377.5
Populus trichocarpa	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Ptrichocarpa/	434.1
Salix purpurea	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Spurpurea	475.5

Table S2. Consensus sequences of Salicaceae SINE families.

SINE Family	Species[a]	Sequence	
		Consensus (5' → 3')	Length [bp]
SaliS-I	*Populus trichocarpa*	AACCATCTAGGTGGTGGCCCAGTGGTAAGAGCTTGGGACCAAGAGGTTTGCTCCCTCTGTGGTCT CAGGTTCGAGCCCTGTGGTTGCTCATATGATGGCCACTGGAGGCTTACATGGTCGTTAACTTCAGG GCCCGTGGGATTAGTCGAGGTGCGCGCAAGCTGGCCCGGACACCCACGTTAAACT	186
SaliS-II	*Populus tremula*	AATTTTGAGGGGTGTAGCTTAACTGGTCAGGTTTTAAATTTGTTTTTTAGAGATCACCAGTTCGAGT CTCACAAATCTCAGGGTCACTGGAGACTTACATGGTCGTTAACTTCAGGACCCGTGAGATTAGTCG AGGTACACGCAAGCTGACCCGAACACCCATATTAAT	169
SaliS-III.1	*Salix purpurea*	GTCCCCAAGGGGCGTGGCGTGATGACAAAGAGCTTGAGATCTGCACAGCAGGTCTCGAGTTCGA GTCAGGACGTGCATCTCTTGTAAGAGTCTGGGACAACCGGGGTTTTACTCGCTCACCTGGACCCA CAAAATACGCTTTCCAGGAGGTGAGGTTTCCTCGAATC	167
SaliS-III.2	*Salix purpurea*	GTCCCCGAGGGGGTGGCGTGGTGGCAAAGGCCTTGGGATCTCCACAGCAGGTCCCAGGTTCGAG TCGCAGGCCATCCCCCCCCTTGTAAGAGCCTGGGACAGCCGGGGGTTTACTCATGCCCTGGGCCCA CAAAGTGCGCTTTCCGGGTCATGTGGTTCCCCCGTATCC	168
SaliS-III.3	*Salix purpurea*	GTCCCCGAGGTGGTGGCCTAGCGGCTRGCGCTTGGGTTCTGCTTCGGCAGACCTGGGTTCGAGCC CGGGAAACAACCCCTCCTCGTAAGAGCCTGGGACAGCCGGAGGTTTACGCATGCCCTGGGCCCAC AAAGTGCGCTTTCCGGGTCATGTGGTTCCCCCGTTCCTATATGGATATCC	180
SaliS-IV.1	*Populus euphratica*	GCAATCAAGGTTTTGGCCTAGCGGTGGAAGGGGCTTGTCTCCTTCCACTGCACCTGGGTTCGAGC CTTGGCGTGCACGCCTGTCACCCCCGCGGTGCCTTACATGCCTACTGGGTTTGCAGGATATTCAGT GGGCCGTGGGGATTAGTCGTGGTGCGCGCAAGCTGGCCCGGACACCCCACGTAAAT	187
SaliS-IV.2	*Populus deltoides*	GCACTTGAGGTTGTAGCTCAGTGGTCAAAGGGACTTGTTTCCTTCCTCTGCTCCCGGGTTCGATCC TYTATGTGCACGCCTGTCACCCCCGCGGTGCCTTACCTGCTCACTGGGCTTGCAGGATGTTCAGTG GGCCGGGGGATTAGTCGTGGTGCGCGTAAGCTGGCCCGGACACCCCAGGTTAT	185
SaliS-IV.3	*Populus tremula*	ACACTTGAGGGTCTAGTTTATTGGTCAACTGCAAGGCTTGTCTTTGCGAATGTCCTGGGTTCGATC CTCAAAGTGTMMGCCTGTCACCCCCGCGGTGCCTTACCTGCCTACTGGGCTTGCAGGATGTTCAG TAGGCCCTGAGATTAGTTGTGGTGCGCGTAAGCTGGCCCGGACACCCAGGATCAT	186
SaliS-V	*Populus trichocarpa*	ACACACGGGGCTTGTAGCTCAGTGGCCTTGGCAGGCTTTGCCCTGCCTGAGTGCCCTGGTTCGAG CTCTCGTGTGTACCCAGCACTTGAGGGTTTAACTGCTGTGGTCCATCGTGGACTTGTTCCGCCCCC CTCCCGGGTTCGACCCTCTATGTGCACGCCTGTCACCCCCGCGGTGCCTTACCTGCTCCTGGGCTT GCAGGATGTCCAGTGGGCCGTGGGGAATAGTCGTGGTGCGCGTAAGCTGGCCCGGACACCCCAC GTGAATC	268

[a] where most copies of the SINE family were identified

Table S2. Continued.

SINE Family	Species[a]	Sequence	
		Consensus (5' → 3')	Length [bp]
SaliS-VI.1	*Populus tremuloides*	ACCAACAGGTTGTGTGGCTCAGTGGTTGTTGGGGGGCAGCTCTCCTTCCTTCGAACTCCGGTTCGA GTCCCAGTGGGAGTGGGGCTGGAGAGTTTGTTCCCTTCCTGTCTCTATTGGGTCCTCCCTGTGCGG TATGCCTGTCACCCCCGCGGTGCCTTACCTGCTCACTGGGCTTGCAGGATGTTCAGTGGGCCGTGG GATTAGTCGTGGTGCGCGCAAGCTGGCCCGGACACCCACGTTAAT	242
SaliS-VI.2	*Salix purpurea*	ACCAGCAGGGGTTGTGGCTCAGTGGTTGTTGGGGGGCGCCCTCCTTCATTCGAACATGGTTCGAG TCCCAGTGGGAGTGGGGCTGGAGGGTTCCTCCCCTTCTTCCTGTCTCCCTTGTTCCCTCCCTGCGC GGTATGCCTGTCACCCCCGCGGTGCCTTACATGCTCACTGAGCTTGCAGGATGTTCAGTGGGCCCG GGGAATAGTCGAGGTGCGCGTAAGCTGGCCCGGACACCCCGGTTAT	243
SaliS-VI.3	*Populus euphratica*	ACCAAGCAGCTTGTAGCTCAGTGGCGTAAGGCGCTGCTCGCCTTCTTTCGAACTTCGGTTCGAGT CCCAGTGGGAGTGGGGCTGGCGAGTGGTTTCCTTCCTGTTTCTGCTGGCTCCTCTCTGTGTGGTAC GCCTGTCACCCCCGCGGTGCCTTACCTGTTCACTGAGGCTTGCAGGATGTTCAGTGGACCGTGGG ATTAGTCGTGGTGCGCGTAAGCTGGCCCGGACACCCACGTTAAT	240
SaliS-VII.1	*Populus deltoides*	ACCCAAGAGGTCCTGGCGGAGCGGTTAGGCGCGCTCTCGTCGCTTACGAGGTTGGGGGTTCGAC CCTTTTCTTCGTCTGCAGCAGGACGCTTGGGGAGGCCTTGCCACCCGGGCCGAGGGATTAGTCTG GGCCAGCGCTTGGAATACCTTGKTTTGAC	158
SaliS-VII.2	*Salix purpurea*	ACCCAAGGGGTCCTGGCGTGAGTGGTGAGGGCGCTCTCGTCCCTTAAGAGAGGTCAGGGGTTCA ATCCCTACTCTTGTATGGAGCTGGCCATTTGGGGAGCACTTTCACCCCTTCGGGGCCCACCCGGTG CGAACGTGGATTAGTCTGGACCAGTGTCTAGGACACCGCGTGGTTTATACC	181
SaliS-VII.3	*Salix purpurea*	ACCCATGAGGTCTCGGCAGAGCGGTTAGGCGCGCTCCTGCCACTGCCGAGGTTGGGGGTTCGAC CCTCTTCCTCGTCAGTAAAGAACCTCATGGGGAGACCTTGCCACCCCGGTCGAGGGATTAGTCTG GGCCGAAGGCCTGGGATACCCTGGTTTGACC	160
SaliS-VII.4	*Salix purpurea*	ACCCATGGGGTCCTGGTCGAGGTGGTAAGGGCAGCCTCGGACCGCTTACGAGGTTGGGGGTTCG ACCCCCTTCTTCGTCTGCAGCAAGGCACATGGGGAGCGCCACCCACGAAGCGCGAAAGGGATTA GTCTGGGCCCTCTGGGTCCAGGATACCTTGGTTATACC	166
SaliS-VIII	*Salix purpurea*	AACCCCAGGGGGTTGGCCTGGCGGTGGAGGCCTGGGGCTCGTGGGTGTGCTCCCCATGAGGTCT CAGGTTCGAATCCGCTCAGGTGCAAACAATTCCTTGGGGCCATCGGACTTGGGCGAAGCCCTTGA CTTAACCGTGGTGCACTTGTGGGAAACATGCTTGCCGAGGCCTTGTGCACCCCCGGGATTAGTCA GGCCCAGCGGCCTGGATACCCGTGGTTAGCTTCGGCTTATCC	236
SaliS-IX	*Salix purpurea*	ACCAACGGGGGCGTGGGTGGACTGGTAGGGGGTCCTCCCAGCTTAACCAAGGTCTCGAGTTCGA GCCTTGGGTATGCAGCTGCTTTAAAACTCGCTTGGGAGAGCTTTGCCGCCCTTAAGGGTCCTACCC GGCTCGAATCCGGATTAGTCGGCAGTGGCCGGATTACCGGATGGTTTCTACC	182

[a] where most copies of the SINE family were identified

Table S2. Continued.

SINE Family	Species[a]	Sequence	
		Consensus (5' → 3')	Length [bp]
SaliS-X	*Salix purpurea*	GACTCCAAGGAGTTGGCCCAACGGTGGAAGCTTGGGACCTGCGTGGGGGTACTCCTCCCAGGTC CTGGGTTCGAAACCTGAGGGAGCAAACTTCTTCTTGGAGTCACCGTCCCCGAGGTGGTGGCCCA GCCCCTCCTCGTAAGAGCCTGGGACAGCCGGAGGTTTACGCATGCCCTGGGCCCACAAAGTGCGC TTTCCGGGTCATGTGGTTCCCCCGTTCCGCAAGGATATCC	233
SaliS-XI	*Salix purpurea*	GTCCCCGAGGGTGTGGTGTAGCGGAAAGWGCTTGGGAGTGGCATCGSCASACCCGGGTTCGAGC CTCTGTATNCCYCCTCGTAAGAGCCTGGGACAGCCGGGGGTTTTAACCGCTCNNCTGGGCCCACA AAGTGCGCTTTCCGGGGAGTGGGGTTTCCCTCGTAAGRATCGATCC	175

[a] where most copies of the SINE family were identified

Table S3. Primers used for synthesis of the Salicaceae SINE probe for fluorescent *in situ* hybridization.

SINE Family	Primer	Amplicon [bp]	Nucleotide Position [bp][a]	Identity [%][a]
SaliS-I	for ATATGATGGCCACTGGAGGC rev CGCGCACCTCGACTAATCCC	66	92 - 157	92.8

[a] regarding consensus sequence

Table S4. Percentage of *P. trichocarpa* Salicaceae SINEs in genes.

SINE Family	CDS	UTR	Intron	<=500 bp	<=1000 bp	<=5000 bp	>5000 bp	Intergenic	Total
SaliS-I	15	9	60	258	292	962	309	1821	1905
SaliS-II	3	8	43	158	200	613	188	1159	1213
SaliS-III.1	2	1	13	34	32	131	60	257	273
SaliS-IV.1	1	1	15	45	58	181	51	335	352
SaliS-IV.2	2	0	3	11	7	37	15	70	75
SaliS-V	1	0	1	9	7	14	1	31	33
SaliS-VI.1	0	0	0	4	2	4	3	13	13
SaliS-VII.3	0	0	0	0	0	1	0	1	1

Table S5. Average similarity of SaliS full-length copies to the species-specific consensus sequences.

SINE Family	*Salix purpurea*	*Populus tremuloides*	*Populus deltoides*	*Populus trichocarpa*	*Populus tremula*	*Populus euphratica*
SaliS-I	70	76	79	88	75	87
SaliS-II	71	72	75	78	71	69
SaliS-III.1	72	81	86	86	79	74
SaliS-III.2	71	-	-	-	-	-
SaliS-III.3	81	-	-	-	-	-
SaliS-IV.1	80	78	-	80	76	82
SaliS-IV.2	-	-	76	97	-	-
SaliS-IV.3	-	79	-	-	84	-
SaliS-V	-	-	94	95	-	-
SaliS-VI.1	-	90	-	-	86	-
SaliS-VI.2	76	-	-	-	-	-
SaliS-VI.3	-	-	-	-	-	84
SaliS-VII.1	90	93	78	91	90	-
SaliS-VII.2	94	97	92	-	-	-
SaliS-VII.3	98	-	-	-	-	-
SaliS-VII.4	100	-	-	-	-	-
SaliS-VIII	93	-	-	-	-	-
SaliS-IX	94	-	-	-	88	-
SaliS-X	93	-	-	-	-	-
SaliS-XI	74	-	-	-	-	-

Table S6. Average values of SINE features, unique for individual copies. A sample of 20 SINE full-length copies with highest similarity to the species-specific consensus sequence was analyzed.

SINE Family	*Salix purpurea*			*Populus tremuloides*			*Populus deltoides*			*Populus trichocarpa*			*Populus tremula*			*Populus euphratica*		
	Tail[a]	TSD[a]	Similarity[b]	Tail[a]	TSD[a]	Similarity[b]	Tail[a]	TSD[a]	Similarity[b]	Tail[a]	TSD[a]	Similarity[b]	Tail[a]	TSD[a]	Similarity[b]	Tail[a]	TSD[a]	Similarity[b]
SaliS-I	8	11	83	10	11	91	12	12	92	12	13	99	12	12	91	13	13	99
SaliS-II	11	10	81	11	11	85	9	12	89	10	12	91	9	10	83	10	9	79
SaliS-III.1	11	11	88	11	12	92	13	12	97	11	14	96	11	13	90	10	12	83
SaliS-III.2	14	13	77	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-III.3	10	12	94	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-IV.1	12	14	98	12	11	93	-	-	-	12	11	88	11	11	89	21	14	96
SaliS-IV.2	-	-	-	-	-	-	13	14	85	17	11	99	-	-	-	-	-	-
SaliS-IV.3	-	-	-	14	14	84	-	-	-	-	-	-	11	11	88	-	-	-
SaliS-V	-	-	-	-	-	-	13	13	94	15	12	97	-	-	-	-	-	-
SaliS-VI.1	-	-	-	11	11	96	-	-	-	-	-	-	14	13	94	-	-	-
SaliS-VI.2	10	12	89	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-VI.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	14	95
SaliS-VII.1	10	14	91	10	13	93	9	16	85	8	14	96	10	15	90	-	-	-
SaliS-VII.2	12	14	99	10	17	97	9	14	93	9	14	-	-	-	-	-	-	-
SaliS-VII.3	10	15	98	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-VII.4	10	15	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-VIII	11	14	99	-	-	-	-	-	-	-	-	-	9	13	-	-	-	-
SaliS-IX	15	14	100	-	-	-	-	-	-	-	-	-	12	17	88	-	-	-
SaliS-X	8	14	96	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaliS-XI	12	14	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

[a] average length
[b] average similarity of full-length copies

Table S7. Overview of 3' end variants of SaliS families and subfamilies.

SINE Family	Species-specific 3' end variants [a]					
	Salix purpurea	***Populus tremuloides***	***Populus deltoides***	***Populus trichocarpa***	***Populus tremula***	***Populus euphratica***
SaliS-I	GTT GTTAATC* GTTAATAAT GTTAATT	GTT GTTAATC* GTTAATCT GTTAATAAT GTTAATT GTTAAAT GTTAAACT	GTT GTTAATC* GTTAATCT GTTAATAAT GTTAATT GTTAAAT GTTAAACT	GTT GTTAATC* GTTAATCT GTTAATAAT GTTAATT GTTAAAT GTTAAACT	GTT GTTAATC* GTTAATCT GTTAATAAT GTTAATT GTTAAAT GTTAAACT	GTTAATC* GTTAATCT GTTAATT
SaliS-II	GTTCC GTTAATC* GTTATC GTTACC	ATT ATTAATC* ATTAATCT ATTAATT	GTT GTTAATC* GTTAATCT GTTAATT GTTAAACT	GTT GTTAATC* GTTAATCT GTTAATT GTTAAACT	ATT ATTAATC* ATTAATCT ATTAATT ATTAAACT	ATT ATTAATC* ATTAATCT ATTAATTGTT
SaliS-III.1	TCGAATC* TCGAATT TCGAATCT	TCGAATC* TCGAATTC TCGAATCTC	TCGAATC* TCGAATTC TCGAATCTC	TCGAATC* TCGAATTC TCGAATCTC	TCGAATC* TCGAATTC TCGAATCTC	TCGAATC* TCGAATTC TCGAATCTC
SaliS-III.2	TCGAATC* CGTTATCC	-	-	-	-	-
SaliS-III.3	GGATATCC	-	-	-	-	-
SaliS-IV.1	GTT GTTATC GTTACC	GTT GTAAATC* GTTATC GTTACC GTCATC	-	GTT GTAAATC* GTTATC GTTACC	GTT GTAAATC* GTTATC GTTACC GTCATC	GTT GTAAATC*
SaliS-IV.2	-	-	GTAAATC* GTTATC	GTAAATC*	-	-
SaliS-IV.3	-	GTTATC GTTACC	-	-	GTTATC ATCATC	-
SaliS-V	-	GTGAATC*	GTGAATC*	GTGAATC* GTAAATC	-	-
SaliS-VI.1	-	GTTAATC*	-	-	GTTAATC* GTT-ATC	-
SaliS-VI.2	GGTTATC GGTTACC GGTCACC GGTCATC	-	-	-	-	-
SaliS-VI.3	-	-	-	-	-	GTTAATC*

[a] Subgroups of different 3' ends had to account for at least 2 % of the SINE family in the respective species. The terminal conserved triplet, upstream of the respective 3' end is underlined. The most frequent 3' end of SaliS-I to SaliS-VI (AATC) is marked with a star (*), and of SaliS-VII to SaliS-XI (ACC) with a circle (°), respectively.

Table S7. Continued.

SINE Family	Species-specific 3' end variants [a]					
	Salix purpurea	***Populus tremuloides***	***Populus deltoides***	***Populus trichocarpa***	***Populus tremula***	***Populus euphratica***
SaliS-VII.1	TAT TATACC°	TATACC°	TTGACCC CACACCC	TCTACC°	TATACC°	–
SaliS-VII.2	TAT TATACC°	TATACC°	TATACC°	TATACC°	–	–
SaliS-VII.3	TTGACC°	–	–	–	–	–
SaliS-VII.4	TATACC°	–	–	–	–	–
SaliS-VIII	TATCC	–	–	–	TATC	–
SaliS-IX	TCTACC° TATACC°	–	–	–	–	–
SaliS-X	TATACC°	–	–	–	–	–
SaliS-XI	GATCC GATTC	–	–	–	–	–

[a] Subgroups of different 3' ends had to account for at least 2 % of the SINE family in the respective species. The terminal conserved triplet, upstream of the respective 3' end is underlined. The most frequent 3' end of SaliS-I to SaliS-VI (AATC) is marked with a star (*), and of SaliS-VII to SaliS-XI (ACC) with a circle (°), respectively.

Supplemental Information to

Chapter 3.1

Localization of the native East Asian origin of the Pillnitz camellia

Content

Supplemental Figures

OZ OSTSEE-ZEITUNG.de

Greifswald / Greifswald

00:00 Uhr / 21.04.2016

Nordeuropas älteste Kamelie steht im Botanischen Garten

Genetische Studie brachte sensationelle Ergebnisse / Greifswalder Exemplar ist mit drei anderen Pflanzen des 18. Jahrhunderts in Pillnitz, Portugal und Italien verwandt

Greifswald. Im Kalthaus des Botanischen Gartens steht eine Rarität: „Die Greifswalder Kamelie ist eine der ältesten in ganz Europa“, sagt Experte Ehrenfried Weidauer. Über 200 Jahre alt ist diese Camelia japonica. Genau wie drei andere Exemplare in Deutschland, Portugal und Italien stammt sie aus dem 18. Jahrhundert und wurde aus Stecklingen der gleichen Mutterpflanze gezogen. Das haben genetische Untersuchungen der Technischen Universität Dresden ergeben. Wahrscheinlich ist die Greifswalder Kamelie sogar die älteste in Nordeuropa.

Thoralf Weiß und Ingrid Handt haben viel Wissenswertes über die Kamelie im Botanischen Garten an der Münterstraße herausgefunden. Quelle: Peter Binder

Wie Thoralf Weiß vom Botanischen Garten informiert, importierte der Botanikprofessor Johann Quistorp 1791 eine Kamelie mit weiteren 72 Pflanzen aus England. In den leider nur bis 1802 und dann wieder ab 1960 geführten Bestandsverzeichnissen ist sie regelmäßig aufgeführt. Auch in einem Herbarium von 1873 findet sich ein Beleg für die derzeit von einem roten Blütenmeer umgebene Pflanze. Das Porzellanschild für die Kamelie soll um 1900 entstanden sein.

Figure S1. Newspaper article about the Greifswald camellia taken from the ‘Ostsee-Zeitung’ (2016).

Im 18. Jahrhundert stand die Pflanze in einem Gewächshaus zwischen Unihauptgebäude und Wall, dem damaligen Botanischen Garten. In ihrem Pflanzkübel wurde sie dann Ende des 19. Jahrhunderts zum heutigen Standort des Botanischen Gartens an der Münterstraße gebracht.

Regelmäßig wird die Kamelie beschnitten. Ab dem 11. Mai kann die inklusive Kübel 3,50 Meter hohe botanische Rarität im Freilandbereich an der Münterstraße wieder besichtigt werden — leider ist sie dann schon verblüht. „Wir bräuchten ein bewegliches Haus wie es im Pillnitzer Schlossgarten vorhanden ist“, meint der Botanikprofessor und Gartendirektor Martin Schnittler.

Dort steht die wohl berühmteste deutsche Kamelie. Sie ist mit etwa neun Metern Höhe und elf Metern Durchmessern das größte Exemplar nördlich der Alpen und wirbt wie die portugiesische und italienische Konkurrenz gern mit dem Superlativ „Älteste Kamelie Europas“. Bisher hieß es immer, dass die Pflanze aus Japan über England nach Pillnitz gelangte. Angeblich hatte Carl Peter Thunberg vier Exemplare von einer Asienreise mitgebracht. Der schwedische Naturforscher gilt als Pionier der Erforschung der japanischen Pflanzenwelt in der Neuzeit. Christian Striefler, Geschäftsführer des Staatsbetriebs der Staatlichen Schlösser, Burgen und Gärten Sachsens, wollte es genauer wissen. Er beauftragte die Untersuchung. „Herr Striefler war hier und hat uns um eine Probe gebeten“, erzählt Weiß.

Die Thunberg-Legende ist nun widerlegt. Die Kamelien stammen auch nicht aus Japan, die Herkunft bleibt noch unbekannt. Die Untersuchung von Proben der Exemplare aus Caserta bei Neapel, aus Campo Bello bei Porto, aus Pillnitz, aus Greifswald und aus Sammlungen zeigen: Das Erbgut der portugiesischen, der italienischen und der beiden deutschen Kamelien ist identisch.

Allerdings steht im Botanischen Garten Greifswald nicht mehr das ursprüngliche Exemplar. Kamelien lassen sich durch Stecklinge vermehren, das war auch hier der Fall, die Pflanzen werden sozusagen geklont. Es wird vermutet, dass die seit 2008 im Kalthaus stehende Pflanze etwa 100 Jahre alt ist. Die Bedeutung der Kamelie liege daher nicht im Alter, sondern in ihrer genetischen Reserve, schätzt die Technische Leiterin des Gartens, Ingrid Handt, ein.

Daten zur Kamelie

1692wurde eine Kamelie erstmals erwähnt.**1777bot Conrad Lodiges** ungefüllte Kamelien, wie sie in Pillnitz und Greifswald stehen, zum Verkauf an. Viel spricht dafür, dass das Greifswalder Exemplar von ihm stammt. Sicher ist der Import aus England. Die Kamelie wechselte im Laufe der Jahre mehrfach den Standort und wurde zuletzt 2008 beim Umzug ins jetzige Haus stark beschnitten.

Von Eckhard Oberdörfer

Figure S1. Continued.

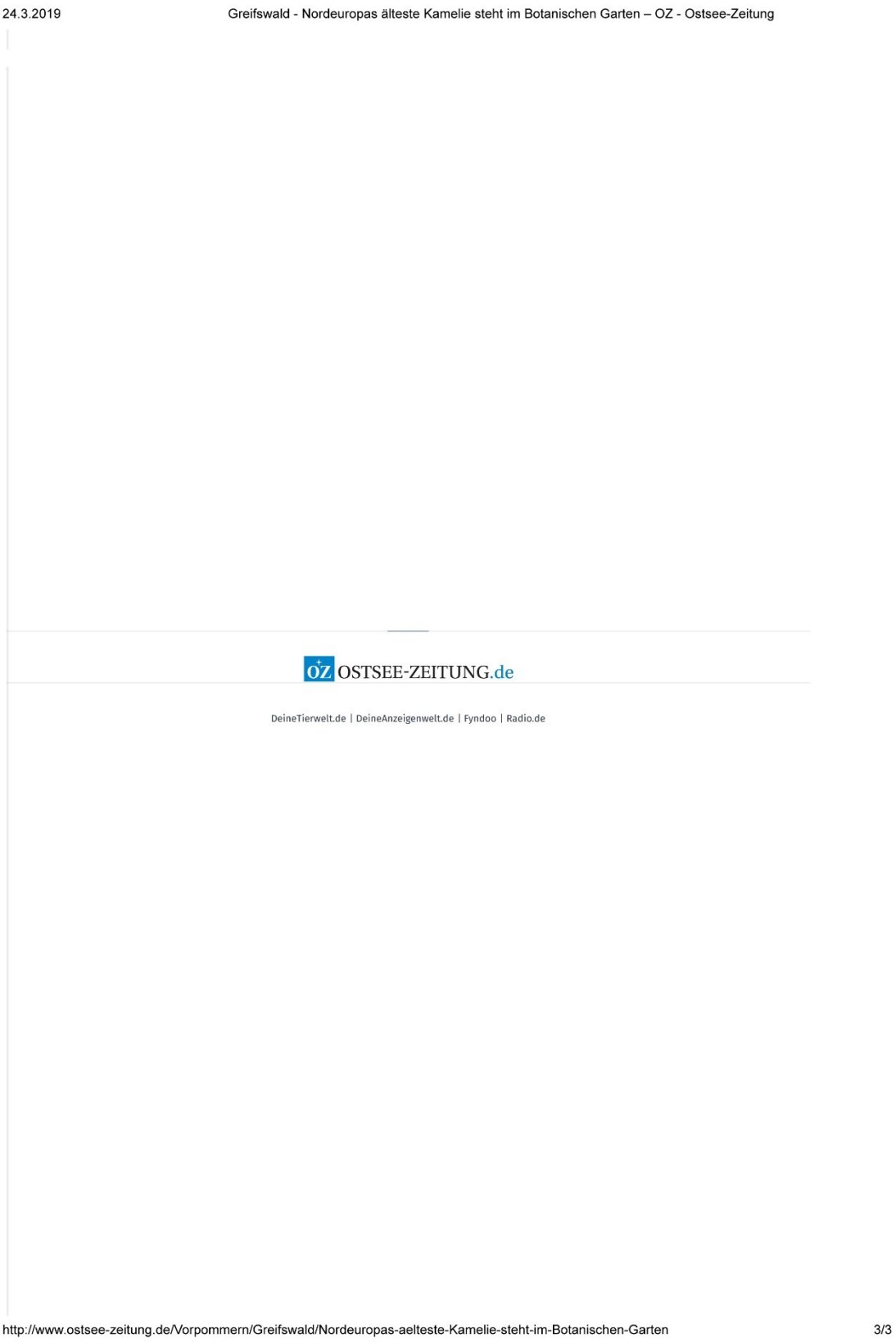


Figure S1. Continued.

Supplemental Information to

Chapter 3.3

Evaluation of the genetic composition of *Larix* hybrids (*Larix* × *eurolepis*) for the targeted identification of profitable phenotypes

Content

Supplemental Figures

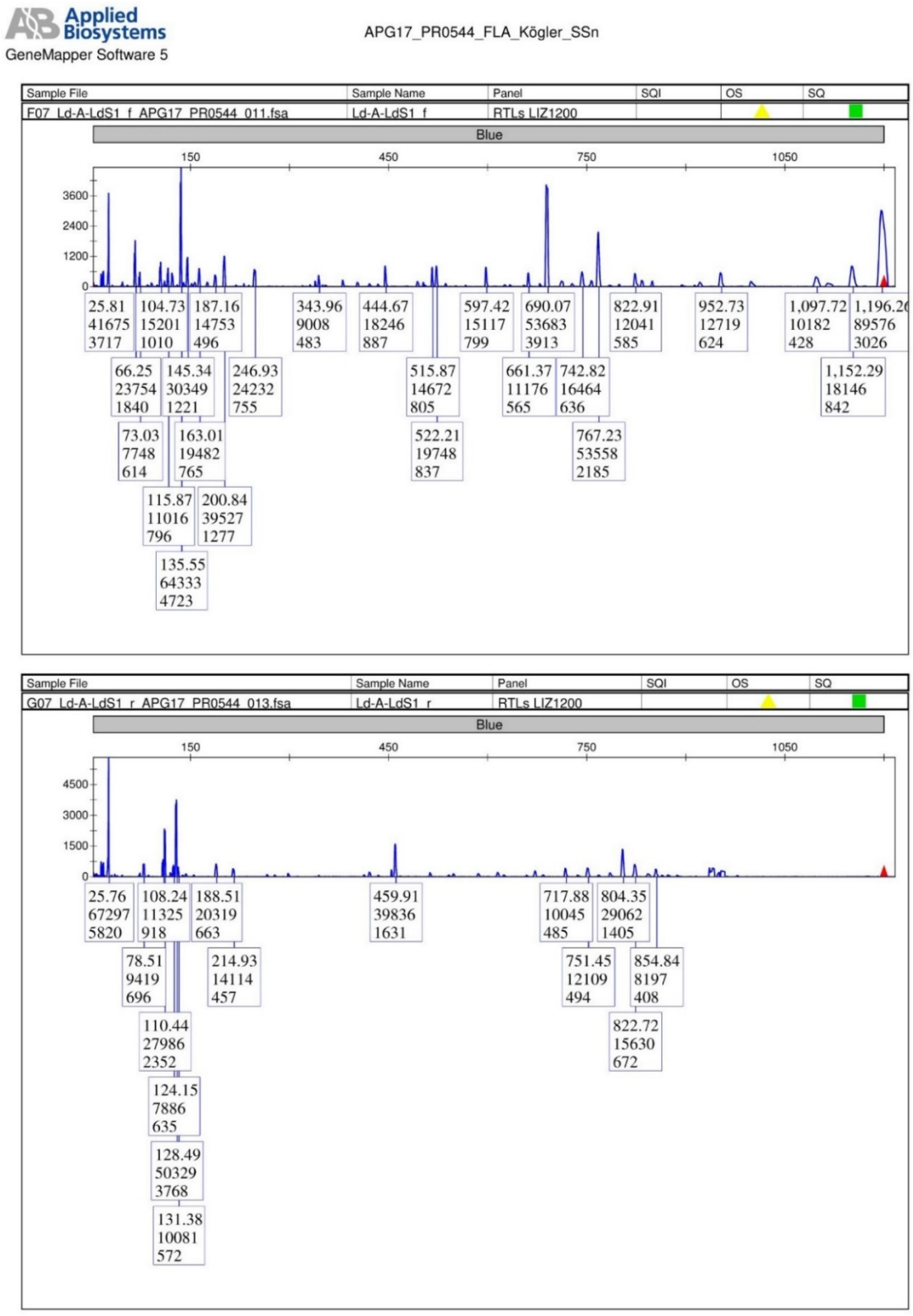


Figure S1. Electropherograms of ISRAP analysis for *L. decidua* 'Tharandt' using the SINE-derived primer LdS-II_*for* (above) and LdS-II_*rev* (below).

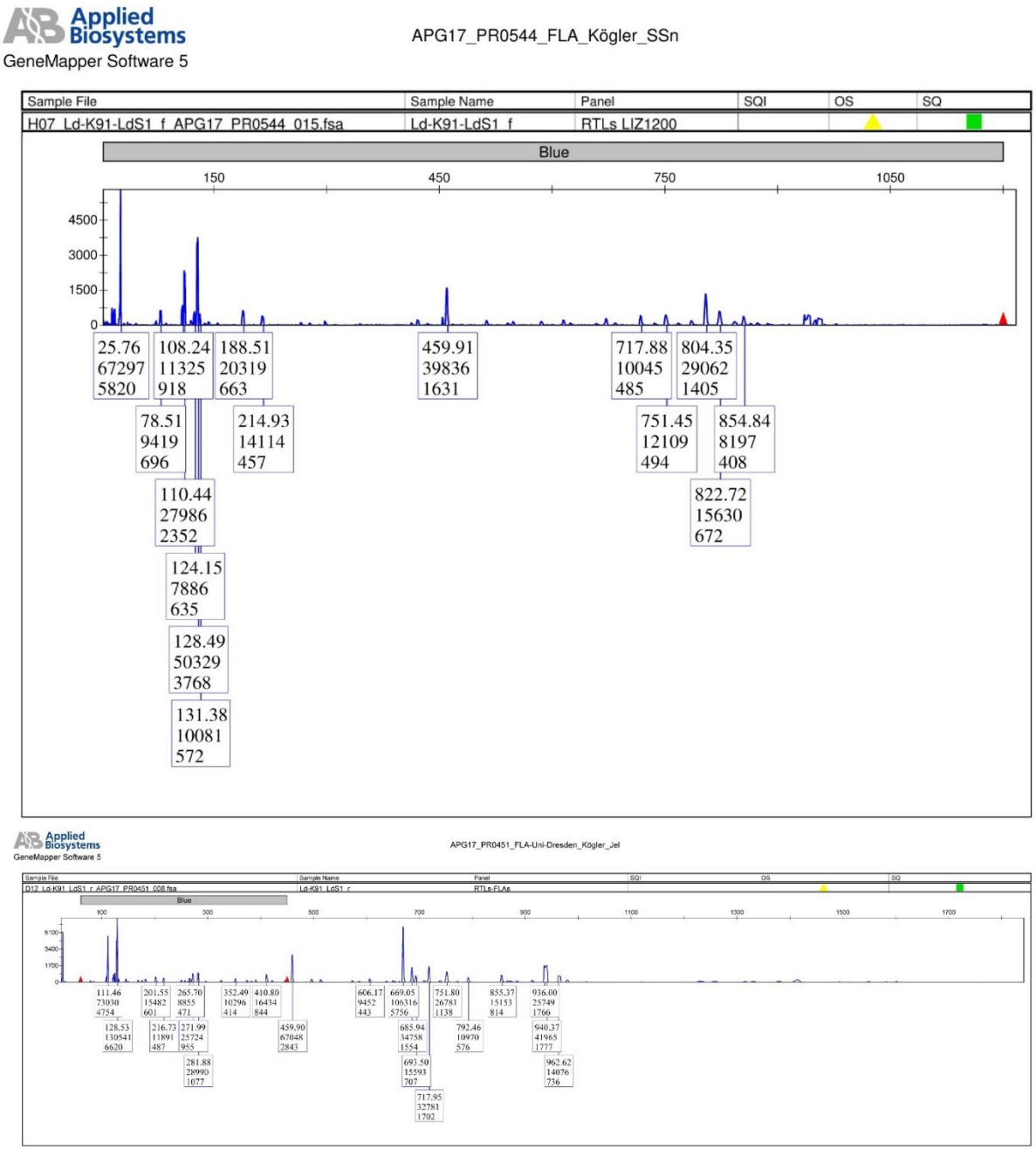


Figure S2. Electropherograms of ISRAP analysis for *L. decidua* 'b-no.91' using the SINE-derived primer LdS-II_*for* (above) and LdS-II_*rev* (below).

Supplemental Tables

Table S1. Fragment length analysis (FLA) peak tables of *L. decidua* genotype comparison using ISRAP.

Peak class	LdS-II_*for*	
	Tharandt	**b-no.91**
Size 1	25.81	25.86
Height 1	3717	5377
Peak Area 1	41675.0	62036.0
Size 2	66.25	65.88
Height 2	1840	2605
Peak Area 2	23754.0	32874.0
Size 3	73.03	
Height 3	614	
Peak Area 3	7748.0	
Size 4		100.91
Height 4		630
Peak Area 4		10338.0
Size 5	104.73	104.56
Height 5	1010	1608
Peak Area 5	15201.0	26515.0
Size 6	115.87	
Height 6	796	
Peak Area 6	11016.0	
Size 7		126.28
Height 7		539
Peak Area 7		8351.0
Size 8	135.55	135.71
Height 8	4723	1897
Peak Area 8	64333.0	31072.0
Size 9		138.59
Height 9		489
Peak Area 9		8423.0
Size 10	145.34	145.39
Height 10	1221	516
Peak Area 10	30349.0	10562.0
Size 11	163.01	
Height 11	765	
Peak Area 11	19482.0	
Size 12	187.16	
Height 12	496	
Peak Area 12	14753.0	

Peak class	LdS-II_*rev*	
	Tharandt	**b-no.91**
Size 1	25.76	
Height 1	5820	
Peak Area 1	67297.0	
Size 2	78.51	
Height 2	696	
Peak Area 2	9419.0	
Size 3	108.24	
Height 3	918	
Peak Area 3	11325.0	
Size 4	110.44	111.46
Height 4	2352	4754
Peak Area 4	27986.0	73030
Size 5	124.15	
Height 5	635	
Peak Area 5	7886.0	
Size 6	128.49	128.53
Height 6	3768	6620
Peak Area 6	50329.0	130541
Size 7	131.38	
Height 7	572	
Peak Area 7	10081.0	
Size 8	188.51	
Height 8	663	
Peak Area 8	20319.0	
Size 9		201.55
Height 9		601
Peak Area 9		15482
Size 10	214.93	
Height 10	457	
Peak Area 10	14114.0	
Size 11		216.73
Height 11		487
Peak Area 11		11891
Size 12		265.7
Height 12		471
Peak Area 12		8855

Table S1. Continued.

Peak class	LdS-II *for*	
	Tharandt	**b-no.91**
Size 13	200.84	
Height 13	1277	
Peak Area 13	39527.0	
Size 14	246.93	
Height 14	755	
Peak Area 14	24232.0	
Size 15		338.57
Height 15		556
Peak Area 15		9241.0
Size 16	343.96	343.54
Height 16	483	786
Peak Area 16	9008.0	15655.0
Size 17	444.67	
Height 17	887	
Peak Area 17	18246.0	
Size 18	515.87	
Height 18	805	
Peak Area 18	14672.0	
Size 19	522.21	522.73
Height 19	837	726
Peak Area 19	19748.0	13491.0
Size 20	597.42	
Height 20	799	
Peak Area 20	15117.0	
Size 21	661.37	
Height 21	565	
Peak Area 21	11176.0	
Size 22	690.07	690.44
Height 22	3913	4918
Peak Area 22	53683.0	75266.0
Size 23	742.82	742.5
Height 23	636	712
Peak Area 23	16464.0	19198.0
Size 24	767.23	
Height 24	2185	
Peak Area 24	53558.0	

Peak class	LdS-II *rev*	
	Tharandt	**b-no.91**
Size 13		271.99
Height 13		955
Peak Area 13		25724
Size 14		281.88
Height 14		1077
Peak Area 14		28990
Size 15		352.49
Height 15		414
Peak Area 15		10296
Size 16		410.8
Height 16		844
Peak Area 16		16434
Size 17	459.91	459.9
Height 17	1631	2843
Peak Area 17	39836.0	67048
Size 18		606.17
Height 18		443
Peak Area 18		9452
Size 19		669.05
Height 19		5756
Peak Area 19		106316
Size 20		685.94
Height 20		1554
Peak Area 20		34758
Size 21		693.5
Height 21		707
Peak Area 21		15593
Size 22	717.88	717.95
Height 22	485	1702
Peak Area 22	10045.0	32781
Size 23	751.45	751.8
Height 23	494	1138
Peak Area 23	12109.0	26781
Size 24		792.46
Height 24		576
Peak Area 24		10970

Table S1. Continued.

Peak class	LdS-II *for*	
	Tharandt	**b-no.91**
Size 25	822.91	
Height 25	585	
Peak Area 25	12041.0	
Size 26		849.57
Height 26		741
Peak Area 26		15946.0
Size 27		903.01
Height 27		426
Peak Area 27		8528.0
Size 28	952.73	
Height 28	624	
Peak Area 28	12719.0	
Size 29		1062.53
Height 29		3028
Peak Area 29		63532.0
Size 30	1097.72	
Height 30	428	
Peak Area 30	10182.0	
Size 31	1152.29	1151.71
Height 31	842	1011
Peak Area 31	18146.0	22233.0
Size 32	1196.26	1192.96
Height 32	3026	4411
Peak Area 32	89576.0	97111.0
total	**25 peaks**	**18 peaks**

Peak class	LdS-II *rev*	
	Tharandt	**b-no.91**
Size 25	804.35	
Height 25	1405	
Peak Area 25	29062.0	
Size 26	822.72	
Height 26	672	
Peak Area 26	15630.0	
Size 27	854.84	855.37
Height 27	408	814
Peak Area 27	8197.0	15153
Size 28		936
Height 28		1766
Peak Area 28		25749
Size 29		940.37
Height 29		1777
Peak Area 29		41965
Size 30		962.62
Height 30		736
Peak Area 30		14076
total	**15 peaks**	**21 peaks**

List of Abbreviations

(v/v)	Volume per volume
(w/v)	Weigth per volume
µl	Microliter(s)
µM	Micromolar
20mer	Oligomer of 20 nucleotides
2n	Diploid chromosome set
40mer	Oligomer of 40 nucleotides
5S rDNA	5S ribosomal DNA
7SL RNA	Signal recognition particle RNA
A	Adenine
A260/A280	Absorbance ratio 260 nm / 280 nm
A550	Fluorescent label related to Rhodamine 6G and Rhodamine B (Eurofins Genomics)
AFLP	Amplified fragment length polymorphism
AmaS	Amaranthaceae SINE
ATTO550	Fluorescent label related to Rhodamine 6G and Rhodamine B (Eurofins Genomics)
BEP	Bambusoideae, Ehrhartoideae, Pooideae
BLAST	Basic Local Alignment Search Tool
BMEL	Bundesministerium für Ernährung und Landwirtschaft (*German Federal Ministry of Food and Agriculture*)
b-no.91	Breeding number 91
bp	Base pair(s)
BSA	Bovine serum albumin
*Bsu*RI (*Hae*III)	Restriction endonuclease isolated from the species *Bacillus subtilis*
C	Cytosine
CCD	Charge-Coupled Device
cDNA	Complementary DNA
CjS	*Camellia japonica* SINE
cpDNA	Chloroplast DNA
CTAB	Cetyltrimethylammonium bromide
Cy3	Cyanine3 fluorochrome
DAPI	4′,6-diamidino-2-phenylindole
DArT	Diversity Arrays Technology
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
dNTP	2´-deoxynucleoside 5´-triphosphate
e.g.	Exempli gratia (*for example*)
*Eco*RI	Restriction endonuclease isolated from the species *Escherichia coli*
*Eco*RI-adap-GA	*Eco*RI adapter primer containing the three selective nucleotides guanine and adenine
*Eco*RI-adap-GAC	*Eco*RI adapter primer containing the three selective nucleotides guanine, adenine and cytosine
EDTA	Ethylenediaminetetraacetic acid
EF-01	Eurofins Dye Set 01
EMBL	European Molecular Biology Laboratory
et al.	Et alia (*and others*)
e-value	Expectation value
FASTA	Fast Adaptive Shrinkage Threshold Algorithm
FISH	Fluorescent *in situ* hybridization
FLA	Fragment length analysis
FNR	Fachagentur Nachwachsende Rohstoffe e.V. (*Agency for Renewable Resources e.V.*)
for	Forward
G	Guanine
Gb	Giga base pair(s)
GBS	Genotyping-by-sequencing
GWAS	Genome-wide association studies
IRAP	Inter-retrotransposon amplified polymorphism
ISAP	Inter-SINE amplified polymorphism
ISRAP	Inter-SINE-restriction site amplified polymorphism
ISSR	Inter-simple sequence repeat
kb	Kilo base pair(s)
Lat	Latitude
LdS	*Larix decidua* SINE
LE	Low electroendosmosis
LINE	Long-interspersed nuclear element
Lon	Longitude
LTR	Long terminal repeat
M	Molar
MAFFT	Multiple Alignment using Fast Fourier Transform
Mb	Mega base pair(s)
MEGA	Molecular Evolutionary Genetics Analysis
mg	Milligram

min	Minute(s)
MITE	Miniature Inverted-repeat Transposable Element
ml	Milliliter(s)
mM	Millimolar
mn	*P. maximowiczii* × *P. nigra*
*Mse*I	Restriction endonuclease isolated from *Micrococcus* species
mt	*P. maximowiczii* × *P. trichocarpa*
mtDNA	Mitochondrial DNA
MUSCLE	Multiple Sequence Comparison by Log-Expectation
mya	Million years ago
N	Any nucleotide; selective nucleotides at the 3' end of an *Eco*RI adapter primer
ng	Nanogram(s)
NGS	Next generation sequencing
ORF	Open reading frame
PACC	Panicoideae, Arundinoideae, Centothecoideae, Chloridoideae
PAM	Point Accepted Mutation
PBS	Primer binding site
PCR	Polymerase chain reaction
PinS	Pinaceae SINE
PKS	Pillnitzer Kamelie Sämling (*Pillnitz camellia seedling*)
PoaS	Poaceae SINE
Pol III	RNA Polymerase III
PPT	Polypurine tract
*Pst*I	Restriction endonuclease isolated from *Providencia stuartii*
PtS	*Populus tremula* SINE
QTL	Quantitative trait locus
RAPD	Random amplified polymorphic DNA
rc	Reverse complementary
REMAP	Retrotransposon-microsatellite amplified polymorphism
rev	Reverse
RFLP	Restriction fragment length polymorphism
rfu	Relative fluorescence unit(s)
RNA	Ribonucleic acid
RNAseq	RNA sequencing
rpm	Revolutions per minute
rRNA	Ribosomal RNA
RT	Reverse transcriptase
s	Second(s)
SaliS	Salicaceae SINE
SCAR	Sequence characterized amplified regions
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
SolS	Solanaceae SINE
SRC	Short rotation coppice
S-SAP	Sequence-specific amplified polymorphism
ssp.	Subspecies
SSR	Simple sequence repeat
STMS	Sequence tagged microsatellite sites
T	Thymine
TAE	Tris base, acetic acid and EDTA
TAIR	The Arabidopsis Information Resource
TE	Transposable element
TheaS	Theaceae SINE
TIGR	The Institute for Genomic Research
TIR	Terminal inverted repeat
TPRT	Target-primed reverse transcription
TRIM	Terminal-repeat retrotransposons in miniature
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
TSD	Target site duplication
TU	University of Technology
U	Unit(s)
UPGMA	Unweighted Pair Group Method with Arithmetic mean
UTR	Untranslated region
vs.	Versus (*unlike*)
WGD	Whole genome duplication

Selbstständigkeitserklärung

Die vorliegende Arbeit wurde am Institut für Botanik am Lehrstuhl für Zell- und Molekularbiologie der Pflanzen der Technischen Universität Dresden unter Betreuung von Prof. Dr. Thomas Schmidt angefertigt.

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel genutzt habe. Die aus fremden Quellen wörtlich oder inhaltlich übernommenen Informationen sind als solche gekennzeichnet.
Ich versichere zudem, dass die vorliegende Dissertation weder in dieser noch in ähnlicher Form in einem anderen Promotionsverfahren eingereicht wurde.

Ich richtete mich nach der Promotionsordnung des Bereiches Mathematik und Naturwissenschaften der Technischen Universität Dresden vom 23. Februar 2011.

Dresden, den 5. September 2019

_____
Anja Kögler