Revealing the past:

The potential of a novel small nucleolar RNA (snoRNA) marker system for studying plant evolution

Gerald Hochschartner

A thesis submitted to theUniversity of St Andrews for the degree of Doctor of Philosophy

School of Biology
University of St Andrews
September 2010

In loving memory of my mother who passed from this life during my time in St. Andrews

CONTENTS

Declaration

I, Gerald Hochschartner, hereby certify that this thesis, which is approximately 50,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

September 2010, Gerald Hochschartner

Statement

I was admitted as a research student in Oktober, 2006 and as a candidate for the degree of PhD in October, 2007; the higher study for which this is a record was carried out in the University of St Andrews between 2006 and 2010.

September 2010, Gerald Hochschartner

Certificate

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

September 2010, Richard J. Abbott

Copyright

September 2010, Gerald Hochschartner

September 2010, Richard J. Abbott

Acknowledgements

Thanks are owed to many people without whom I would not have been able to write this thesis.

Firstly, I would like to thank my supervisors, Richard Abbott and John Brown for their constant support, helpful advice, encouragement and guidance during the course of my PhD. Particular thanks also go to Joanne Russell for her supervision in the lab and valuable help in conducting the research for this thesis. I would also like to thank David Forbes for his assistance with DNA extractions and his expertise in the lab, Adrian Brennan and Charles Coyle for constructive discussions and good company, and Tom Meagher and Jeff Graves for their valuable comments during our few meetings. I am also most grateful to my colleagues at SCRI, especially Alan Booth who had always an open ear for my needs in the lab and his wife Claire who ran my samples on the sequencer. Special thanks go to Heather Ross for her save lifts to SCRI and back to St. Andrews.

I would also like to thank my many colleagues and friends in St. Andrews for making my stay in this beautiful town so enjoyable. Special thanks to Clint, Luca and all the people from the Harold Mitchell building.

Many thanks are owed to my family and friends in Austria for keeping contact and providing support and encouragement throughout my stay in St. Andrews. Special thanks go to my sister Christine and my friends Hansi, Martin, Andi, Wolfgang and Paul who came to Scotland and accompanied me on various trips around this beautiful country.

I am also very grateful to my new parents in law, Margareta and Wolfgang for their overwhelming support and help during the final year of my PhD; without them writing this thesis would have been much more difficult.

I would also like to thank the University of St. Andrews and the Scottish Crop Research Institute for funding the research and financial support.

Finally, I am eternally thankful to my beautiful wife Ricarda who I met and became engaged to in St. Andrews. Throughout my PhD studies and especially during the last stage of the writing process in Austria when I had my up and downs away from the university environment, she was always there for me with her love and patience to provide endless support and encouragement. Thank you.

Abstract

Despite the existence of various molecular marker systems there are still limitations in distinguishing between closely related species based on molecular divergence, especially when hybridization events have occurred in the past. The characterisation of plant small nucleolar RNA (snoRNA) genes and their organisation into multigene clusters provides a potential nuclear marker system which could help in resolving the phylogenetic history of plants and might be applicable in DNA barcoding. Using closely and distantly related *Senecio* species, I investigated a combination of fragment length and sequence variation of snoRNA genes/snoRNA gene clusters to assess the utility of this marker system for barcoding and resolving species relationships.

SnoRNA gene and gene cluster sequences identified in *Arabidopsis thaliana* were used to find homologues in other species and subsequently used for the design of universal primers. Most of the universal primer pairs designed were successful in amplifying snoRNA fragments in most *Senecio* species and fragment length variation between and within species could be detected. Furthermore, the combination of some fragment length datasets produced by different primer pairs enabled the separation of species and the detection of reticulate evolution indicating a high potential of snoRNA gene/gene cluster fragment length polymorphisms (SRFLPs) for phylogenetic reconstructions in *Senecio* and other plant genera.

Most of the examined gene clusters showed a similar gene order in *Senecio* and *Arabidopsis*. However, the majority of these clusters appeared to exhibit more copies in *Senecio*, some of which were distinguishable by a combined sequencing/fragment profiling approach, and shown to be putative single copy regions with the potential to be used as co-dominant markers. However, a high number of paralogues and possible differences in copy number between species excludes these regions from being used in DNA barcoding. This is because specific primers would have to be developed for specific copies which would preclude development of a universal application for barcoding.

None of the regions showed enough sequence variation to delimit distinctly closely related *Senecio* species and were therefore also considered to be unsuitable for DNA barcoding. Although most snoRNA genes and gene clusters might be inapplicable

for DNA barcoding, they are likely to be valuable for phylogenetic studies of species groups, genera and families. On this scale, specific primers might act universally and the number of paralogous copies is likely to be equal across the species group of interest.

Chapter 1: Introduction

Over the last five decades molecular markers have greatly contributed to an improved understanding of biological patterns and processes. Progress in molecular biology has led to the development of a wide range of markers for studies of evolution and related phenomena including the analysis of population genetic structure, speciation, phylogenetic relationships and DNA barcoding. Molecular markers can be divided into three categories: alloenzymes, DNA sequence polymorphisms and DNA repeat variation (Schlotterer, 2004). Nowadays, the method of choice, especially for phylogenetic analysis and DNA barcoding is DNA sequencing (e.g. Chapman *et al.*, 2007; Hollingsworth *et al.*, 2009a; Hollingsworth *et al.*, 2009b). While phylogenetic studies attempt to reconstruct ancestral relationships between species (Schlotterer, 2004), DNA barcoding is concerned with species identification (e.g. Hebert *et al.*, 2003). However, the requirements of markers for both forms of analysis are very similar and markers useful for DNA barcoding are most likely applicable for phylogenetic reconstruction. A major aim of the present thesis is to determine whether small nucleolar RNA (snoRNA) might have the potential to be developed as a marker system suitable for both DNA barcoding and phylogenetic analysis in plants.

1.1 DNA-Barcoding

DNA-barcoding is currently of wide interest in taxonomy, biodiversity studies and evolutionary biology (Kadereit, 1984; Chase *et al.*, 2005; Kress *et al.*, 2005; Monaghan *et al.*, 2005; Savolainen *et al.*, 2005; Kane & Cronk, 2008). It may be used for species identification, the discovery of new species, and the construction of biodiversity inventories in the field (Savolainen *et al.*, 2005). A two step approach, however, is necessary to resolve difficult taxonomic problems, caused for example by reticulate evolution and incomplete lineage sorting. In the first step only a single sequence is used. If the result is satisfactory in terms of taxon identification then no further investigation is necessary, but if not, multilocus sequencing (i.e. sequencing more than one region of the

genome) will hopefully provide taxonomic clarity in problematic groups (Chase *et al.*, 2005).

1.1.1 Species identification and classification

Although molecular techniques, particularly DNA-sequencing, are now widely used in biogeographical, ecological and evolutionary studies, taxonomic studies, except those of bacteria, still rely heavily on morphological characters. For every new species described, a type specimen has to be deposited in a major museum/herbarium collection that can be accessed for inspection and analysis. It is argued that a more user friendly taxonomic system might be introduced, which allows a more rapid identification and classification of species (Chase *et al.*, 2005). Thus, a 'temporary' taxonomy of the group of interest could be obtained rapidly using molecular techniques such as DNA-sequencing. Afterwards, these data (stored in databases) could be used by taxonomic specialists and, together with other data (e.g. morphological) a proper taxonomy, sometimes referred to as 'reversed taxonomy', for the group could be obtained (Markmann & Tautz, 2005; Monaghan *et al.*, 2005; Savolainen *et al.*, 2005). In some species groups (e.g. animal species) only one sequence might be enough to obtain a robust 'temporary' taxonomy, although additional sequences may be required when problems arise due to the effects of introgression and/or incomplete lineage sorting.

The major challenge to adopting a bar-coding approach in taxonomy is to find useful sequences (markers), which distinguish species (and subspecies) and exhibit much lower variation within than between species. An appropriate marker should be short and, thus, ideally available from degraded samples (e.g. from old herbarium specimens), and easy to amplify in almost all species-groups. In addition, the marker of choice should (i) lack divergent paralogues, otherwise cloning of multiple copies will be necessary, (ii) not cause secondary structure problems because this could lead to poor results in both the amplification- and the sequence-reaction (Alvarez & Wendel, 2003; Blaxter, 2004; Kress *et al.*, 2005), and (iii) be generated using a universal primer pair and be accessible to bidirectional sequencing without much manual editing of sequence traces (Hollingsworth *et al.*, 2009b).

1.2 Existing 'universal' markers

1.2.1 Cytoplasmic markers

One marker that has been successfully used in phylogenetic studies and DNA-barcoding of animals, as well as some fungi and some algal groups (Chase *et al.*, 2005; Saunders, 2005) is the mitochondrial subunit 1 of the cytochrome b gene (*cox*1). In most plant groups, however, the mitochondrial genome is not useful for barcoding because of its low sequence-variation and rapid change in structure (Adams & Palmer, 2003). Therefore, the chloroplast genome has been examined to find appropriate regions that might be used in barcoding. As various chloroplast (cp) DNA-regions (e.g. *rbc*L, *mat*K and *ndh*F exons; *trn*L-F and *trn*H-*psb*A intergenic spacers) are widely used for plant systematic and biogeographic studies (Shaw et al., 2005), these have been tested for their appropriateness in barcoding. Some of these (e.g. *mat*K, *rbc*L, *trn*L-F, *trn*H-*psb*A) appear to be useful in discriminating between species (Kadereit, 1984; Chase *et al.*, 2005; Kress *et al.*, 2005; Kane & Cronk, 2008; Lahaye *et al.*, 2008; Hollingsworth *et al.*, 2009a; Hollingsworth *et al.*, 2009b), but only two (*rbc*L and *mat*K) were recently recommended as components for a standard 2-locus barcode in plants (Hollingsworth *et al.*, 2009b). Although the chloroplast genome shows strikingly high conservation across the plant kingdom, non-photosynthetic plants are the exceptions. The plastid genomes of these species show gene loss, gene retention and accelerated evolutionary rates, particularly in photosynthesis associated regions making most cp markers unsuitable for these taxa (dePamphilis & Palmer, 1990; Bungard, 2004). Additional regions will need to be used to separate taxa in difficult plant groups (Shneyer, 2009; Le Clerc-Blain *et al.*, 2010; Mort *et al.*, 2010), and in those where reticulate evolution has occurred biparentally inherited nuclear markers are likely to be necessary for accurate barcoding (Alvarez & Wendel, 2003; Chase *et al.*, 2005).

1.2.2 Nuclear markers

The lack of nuclear sequence information available for universal primer design and difficulties in identifying orthologues and paralogues have impeded the development of universal nuclear markers (Small *et al.*, 2004). Consequently, the internal transcribed spacer (ITS) (Figure 1.1) of the nuclear ribosomal cistron (18S rDNA-5.8S rDNA-26S rDNA), which consists of ITS1, 5.8S rDNA and ITS2, has been widely used in phylogenetic analysis and, thus, a large number of ITS sequences already exists for plant species.

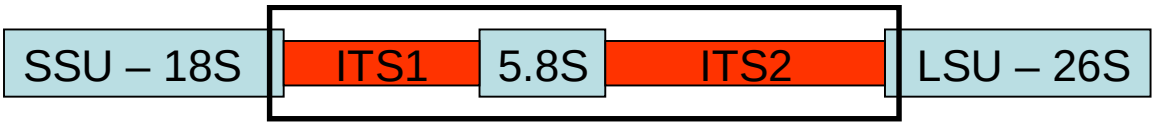


Figure 1.1: The internal transcribed spacer (ITS) region (box) is used in many phylogenetic studies and consists of ITS1, 5.8S rDNA and ITS2. SSU = Small subunit; LSU = Large subunit.

ITS sequences have been used in many phylogenetic studies because they were thought to exhibit biparental inheritance, universality, simplicity, intragenomic uniformity, intergenomic variability and low functional constraints (Baldwin *et al.*, 1995; Alvarez & Wendel, 2003). Biparental inheritance is important for resolving cases of reticulation, hybrid speciation and the parentage of polyploids as shown in some previous studies (Rieseberg *et al.*, 1990; Rieseberg & Soltis, 1991; Baldwin *et al.*, 1995; Wendel *et al.*, 1995). A set of primers has proved useful for amplifying the ITS-region in most fungal and plant taxa, which means that no new primer design is normally necessary for obtaining new sequence information from unknown groups or taxa. As the ITS-region is relatively short in angiosperms (500 to 700 bp; Baldwin *et al.*, 1995) and exists in hundreds to thousands of copies distributed over one or more chromosomes in each plant genome (rDNA arrays), it can be amplified successfully, even from old herbarium samples. Furthermore, in cases where the entire ITS sequence cannot be obtained with one reaction, there is the possibility to amplify the ITS1 and ITS2 sequences separately using internal primers.

In some groups, however, ITS sequences show reduced interspecific variability, particularly in recently diverged taxa such as the *Bulbophyllum lobbii* complex, Orchidaceae (Hochschartner, 2006), and/or lack of intragenomic uniformity due to divergent paralogues as for example observed in *Ophrys* (Gulyas *et al.*, 2005), *Oryza* (Bao *et al.*, 2010), *Cycas* (Xiao *et al.*, 2010) and *Eucalyptus* (Bayly & Ladiges, 2007), and secondary-structure problems. Orthologous and paralogous sequences may merge, basepair changes may be compensated, indel accumulation may cause alignment problems and, thus, might lead to higher levels of homoplasy than exhibited by other DNA sequences. It has also been noted that divergent rDNA copies may undergo a variety of fates after combining them in a single genome due to reticulation (Wendel, 2000). Alvarez and Wendel (2003) discuss three different possible evolutionary fates, which are not mutually exclusive (Figure 1.2). One is that there is a lack of concerted evolution (Figure 1.2-1), causing divergent rDNA copies to remain present in a genome for a long time, such as in the Winteraceae (Suh *et al.*, 1993). In this case, recombination and inter-array exchange does not occur rapidly and therefore mutations can accumulate which leads to independent evolution of divergent rDNA copies. The lack of homogenization might be very informative in discovering both past hybridization and polyploidization events and maternal and paternal progenitor lineages (e.g. in Tragopogon allopolyploids: Soltis *et al.*, 1995). A second possibility is that different rDNA types will remain and recombine to various degrees (Figure 1.2-2). This phenomenon, which might be common in hybrids (Campbell *et al.*, 1997; Barkman & Simpson, 2002) will lead to chimeric rDNA sequences which will be basal to either parent lineage (McDade, 1992). For example, chimeric ITS sequences might consist of an ITS1 from one and an ITS2 from the other parental lineage, and, thus, be composed of different ribotypes, which have undergone genic recombination (Barkman, Simpson, 2002). Such combination of sequences can occur between different functional repeats and also between functional and non-functional repeats. A third possible fate is that only one type of rDNA repeat remains due to concerted evolution (Figure 1.2-3). This sequence could either be an uncontaminated offspring type of one parent lineage or a chimeric composition of both parental ribotypes resulting from intergenomic recombination. Elimination of rDNA repeats could take place by intergenomic recombination following

allopolyploidization and may be bidirectional, so that in one descendant there will remain the sequence of one parent and in an alternative one that of the other parental lineage (Wendel & Cronn, 2003).

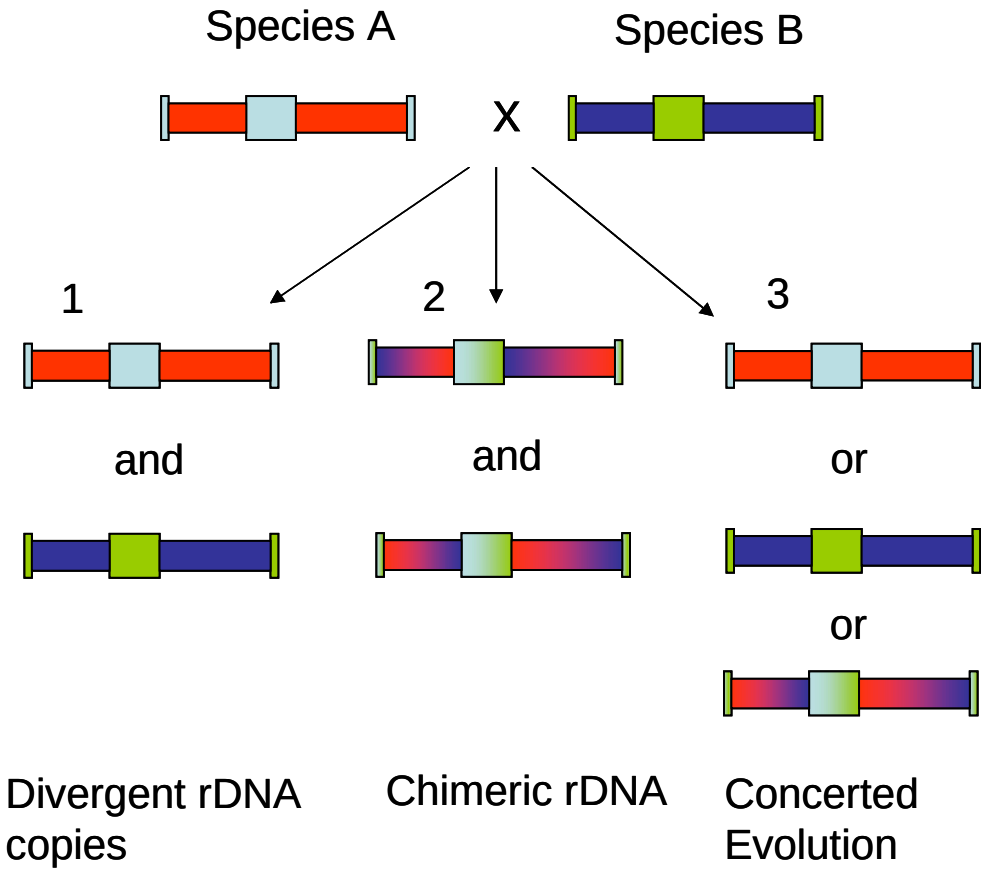


Figure 1.2: The three fates of ITS. Lack of concerted evolution (3) might result in divergent copies of rDNA, either uncontaminated (1) or chimeric (2). Note that many different copies could accumulate over time.

It is now clear that insights into the history of plant species, such as reticulation, hybridization and allopolyploidization, can hardly be obtained by direct sequencing of a single PCR amplification reaction, but rather demands cloning and multiple sequencing. However, as orthologues may be lost in only some taxa, problems of both paralogue comparisons, which increase with ploidy level and the formation of dead or dying repeats (pseudogenes), have to be taken into account. Non-functional repeats or pseudogenes of various ages may persist in the genome and evolve independently and at different rates

relative to functional repeats. Thus, orthologues could merge with paralogues and pseudogenes, respectively, creating chimeric sequences. Such complex patterns of paralogy could lead to erroneous phylogenies (Alvarez & Wendel, 2003).

ITS sequences also do not evolve entirely neutrally because secondary structure (e.g. stem-loop) is important for their function. A high GC-content provides stability of these secondary structures and, thus, compensation of base changes should be frequent at these sites (Mai & Coleman, 1997). Such compensatory base changes can lead to homoplasy. It is also noted that as ITS is not coding for a protein, indel accumulation could cause problems with alignment. Short indels can arise due to DNA replication slippage (Hancock & Vogler, 2000), while longer indels are also common in ITS sequences.

In summary, although ITS variation has been considered to be very suitable for examining phylogenetic relationships, and consequently ITS sequences for a wide range of taxa are available (e.g. Genbank), there are problems in using ITS sequence variation in phylogenetic and barcoding analyses.

More accurate results might be obtained using single or low-copy nuclear genes (Alvarez & Wendel, 2003) possibly provided by a novel snoRNA marker system. SnoRNA genes might be useful for DNA-barcoding and phylogenetic approaches in plants because they (i) might evolve faster than protein-coding genes (no open reading frame, individual function nonessential, only conserved regions for stability and modification function), (ii) are clustered (fixed or unfixed order) and spread over the whole genome, (iii) have short sequences, (iv) have conserved regions which can be used for universal primer design, (v) have a low number of copies which should not be subject to concerted evolution and (vi) might have lower homoplasy compared to ITS.

1.3 Small nucleolar RNA

Small nucleolar RNAs (snoRNAs) belong to a large family of non-coding RNAs and are usually 60 to 300 nucleotides (nt) long. They are mainly involved in the biogenesis of ribosomal RNAs (rRNAs) (Maxwell & Fournier, 1995) and also play a role in post-transcriptional modification of transfer RNAs (tRNAs) (Zemann *et al.*, 2006) and snRNAs (Tycowski *et al.*, 1998; Darzacq *et al.*, 2002; Marker *et al.*, 2002; Chen *et al.*, 2008), and in pre-mRNA splicing (Bachellerie & Cavaille, 1997; Kishore & Stamm, 2006; Nahkuri *et al.*, 2008). Additionally, snoRNAs with unknown targets have been identified and are referred to as orphan snoRNAs (Marker *et al.*, 2002; Chen *et al.*, 2008).

The first small nucleolar RNAs (snoRNA) were discovered in mammalian cells in the late 1960s, but it needed another 20 to 25 years and new experimental methods and technologies to progress considerably in research on snoRNAs. Today we know that snoRNAs are found in eukaryotic cells and also in archaea where they are called sno-like RNAs (sRNA) due to the lack of nucleoli (Maxwell & Fournier, 1995; Makarova & Kramerov, 2007). To date, many different snoRNAs have been identified and named using two different systems of classification. Because they contained a high number of uracil bases, the first small RNAs to be discovered (small nuclear RNAs – snRNAs), which are involved in precursor messenger RNA (pre-mRNA) splicing, were called U1, U2, U4, U5 and U6. Although the majority of snoRNAs are not characterised by high uracil content, newly identified snoRNAs and their orthologues in other organisms have, in turn, received the next free U-number when added to this classification sequence (Busch *et al.*, 1982). However, an exception to this is in yeast where snoRNAs have been assigned different snR-numbers, either depending on their positions in a 2-D polyacrylamide gel (snR1, snR2, etc.) or the estimated number of nucleotides they contain (snR189, snoR190, etc). The first system of classifying snoRNAs is still in use, whereas the latter one has left some gaps in the snR-numbers (Riedel *et al.*, 1986; Thompson *et al.*, 1988; Maxwell & Fournier, 1995). For example, in yeast, snR87 is followed by snR161 (UMASS Amherst yeast snoRNA database: http://people.biochem.umass.edu/sfournier/fournierlab/snornadb/mastertable.php).

1.3.1 Structure and function of snoRNAs

There are two main types of snoRNA which differ in structure and function: the box C/D snoRNAs (Figure 1.3A) are responsible for 2'-O-ribose methylation of ribosomes (Figure 1.3C) and the box H/ACA snoRNAs (Figure 1.3B) for pseudouridylation (Figure 1.3C). Both classes of snoRNAs are associated with proteins forming small nucleolar ribonucleoprotein particles (snoRNPs) where the proteins are necessary for stability of the snoRNP to perform modification functions. SnoRNAs have guide functions in determining the nucleotide for modification via a base-pairing interaction between the snoRNA and target RNA (Brown *et al.*, 2003a).

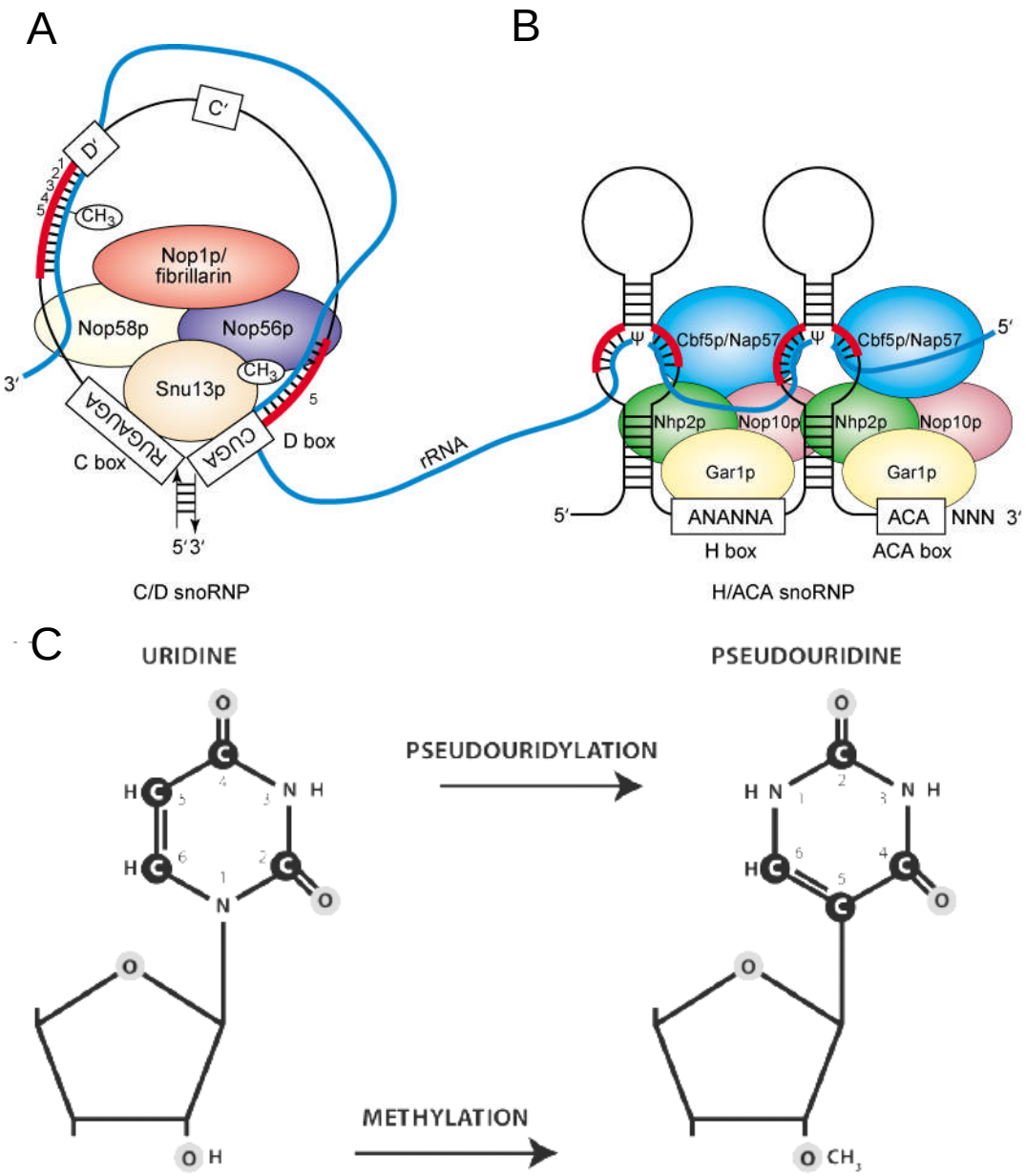


Figure 1.3: The major types of snoRNAs and their associated proteins. A: The box C/D snoRNPs are responsible for the (C) 2'-O-ribose methylation of the ribosomes. Note that an additional Nop1p/fibrillarin and Nop56 protein can be bound to the box C'/D' motif (not shown) leading to structural asymmetry. Furthermore, some box C/D snoRNAs have a loop between the C' and D' box. B: The box H/ACA snoRNPs convert (C) uridine to pseudouridine. Figures are taken from (Brown *et al.*, 2003a) and (Henras *et al.*, 2008).

1.3.2 Box C/D snoRNAs

All box C/D snoRNAs contain the well conserved box-C (at the 5' end) and box-D (at the 3' end), with the consensus sequences RUGAUGA and CUGA, respectively (Figure 1.3A). Normally, the C and D boxes are brought together by short inverted repeats at the 5' and 3' end of the snoRNAs, respectively, which are able to form a terminal stem of 3 to 8 bp in length (Brown *et al.*, 2003a; Nahkuri *et al.*, 2008). Although the stem structure is necessary for snoRNA biogenesis, the primary sequence is not well conserved (Bachellerie *et al.*, 2002) and, interestingly, is partially or fully degraded in some mature snoRNAs (Darzacq & Kiss, 2000). The resulting structure (stem, box-C and D) is called the C/D motif. A closer look at this C/D motif reveals a more complex structure, the so-called kink-turn (K-turn) which acts as a platform for C/D snoRNP proteins (Figure 1.4). Furthermore, the C/D motif is not only necessary for the formation, but also for the stability and nucleolar transportation of the snoRNPs (Samarsky *et al.*, 1998; Henras *et al.*, 2004b; Makarova & Kramerov, 2007). In addition to the highly conserved box C and D elements, box C/D snoRNAs contain more or less degenerate box C' and D' in their centre, which are usually just 3 to 9 nucleotides (nt) apart or brought together by a intermolecular hairpin in cases of greater distance (Kiss-Laszlo *et al.*, 1998; Brown *et al.*, 2001).

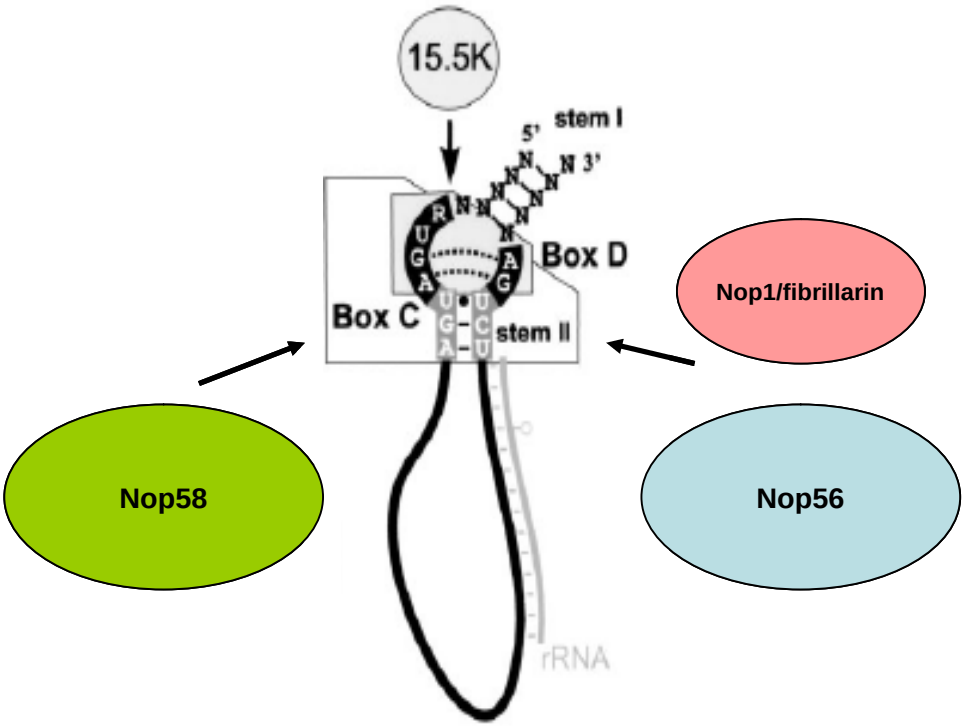


Figure 1.4: Hierarchical assembly of box C/D snoRNPs. After the binding of the 15.5K protein to the internal loop of the C/D motif the remaining box C/D snoRNP proteins can be recruited. Figure is modified from (Watkins *et al.*, 2002)

The guide function of the box C/D snoRNAs resides in one or two RNA antisense elements of between 10 and 21 nucleotides which base-pair with a specific RNA target region (e.g. ribosomal RNA). The antisense sequences are adjacent to and upstream of the box D and/or the internal box D' elements (Figure 1.3A). The nucleotide of the target RNA which is methylated is the fifth residue from the D or the D' box which happens to be at the half turn of the helix (Tollervey, 1996). In eukaryotes, most snoRNAs contain a single guide sequence, but some contain two antisense elements. For instance, in *Arabidopsis* only a quarter of the snoRNAs contain two antisense elements. The opposite is true for archaea where a large fraction of snoRNAs contain two antisense elements. Normally, when two elements are present, they target residues which are either close to each other due to primary (sequence) or structural proximity (Barneche *et al.*, 2001). Interestingly, snoRNAs have been identified with two antisense elements modifying just one possible target RNA (Russell *et al.*, 2006).

The conserved sequence elements were first defined by comparing different U3 snoRNA homologues and were given the letters A to D. It turned out that the boxes A and B could only be found in U3, and boxes C and D were present in one major subset of snoRNAs that were therefore named box C/D snoRNAs.

1.3.3 Box H/ACA snoRNAs

The box H/ACA snoRNAs have shorter conserved motifs (box H and box ACA, respectively) and antisense elements. The box-H (Hinge) element links two hairpin structures and its highly degenerated consensus sequence is ANANNA. The highly conserved box-ACA (consensus sequence ACANNN) is 3 nt away from the 3' end at the base of the second hairpin (Figure 1.3B). This whole structure containing the two hairpins and the two single-stranded boxes is known as the hairpin-hinge-hairpin-tail and is crucial for pseudouridylation efficiency (Ganot *et al.*, 1997a; Ganot *et al.*, 1997b; Bortolin *et al.*, 1999). Similar to the box C/D snoRNAs, the H/ACA snoRNAs contain either one or two antisense elements which are located in the hairpins. The antisense element consists of two separated sequences each of 3 to 10 nt, which lie in an internal loop formed by the secondary structure of its hairpin. The target RNA base-pairs with the antisense elements and leaves two nucleotides unpaired at the top of the internal loop, one of which is targeted for pseudouridylation. Uridine is converted to pseudouridine by a 180° rotation of the Uracil residue around its N_3-C_6 axis, breakage of the C_1-N_1 bond and formation of the new C_1-C_5 bond (Figure 1.3C). The nucleotide for pseudouridylation in the target RNA has a conserved distance to the box-H and/or box-ACA of 14 to 15 nt (Brown *et al.*, 2003a; Makarova & Kramerov, 2007). While archaeal box H/ACA sRNAs might contain a K-turn in their apical hairpin, no such motif was found in box H/ACA snoRNAs of eukaryotes (Rozhdestvensky *et al.*, 2003; Reichow *et al.*, 2007).

1.3.4 5’ and 3’ end modifications of snoRNAs

The 5’ end of snoRNAs is largely determined by mode of expression (see below). Firstly, snoRNAs transcribed from their own promoter contain the 2,2,7-trimethylguanosine (TMG) cap (Figure 1.5B). This is found in most yeast and half of vertebrate RNAs examined and is thought to be formed on the ends of RNA polymerase II transcripts by hypermethylation of the 7-methylguanosine cap (Figure 1.5A) during snoRNP biogenesis (Terns *et al.*, 1995; Terns & Terns, 2002; Watkins *et al.*, 2004).

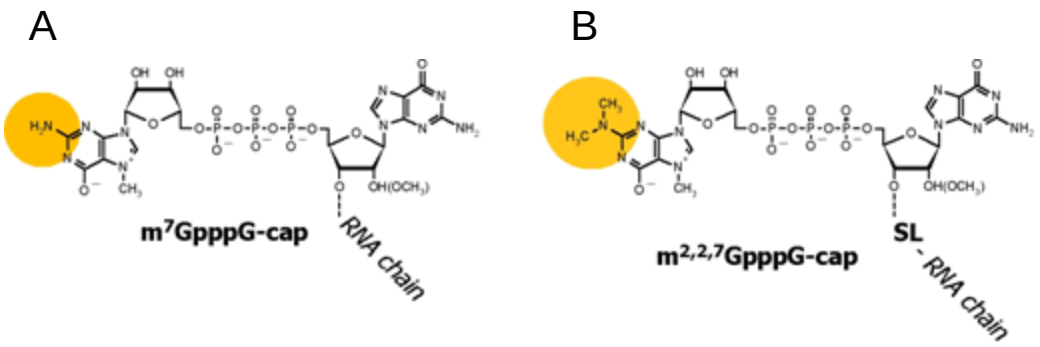


Figure 1.5: Cap structures produced by RNA polymerase II. A: The 7-methylguanosine cap becomes B: hypermethylated to the TMG cap (right) in mature snoRNAs. Hypermethylation site in yellow circles. Figure is taken from (http://www.uchsc.edu/molbio/davisr.htm).

Secondly, a γ-monomethyl phosphate cap (Figure 1.6) was found at the 5’ end of the plant snoRNA U3 (Shimba *et al.*, 1992) and human U6 (Singh & Reddy, 1989) transcribed by RNA polymerase III. Thirdly, a few snoRNAs in yeast and more than half of mammal snoRNAs do not possess a cap at all but instead have an unmodified 5’ monophosphate. This is usually produced as a result of processing pre-snoRNAs located in introns. The 3’ terminus of snoRNAs possesses a simple OH-group (Liu *et al.*, 1992; Reddy *et al.*, 1992; Maxwell & Fournier, 1995; Terns & Terns, 2002).

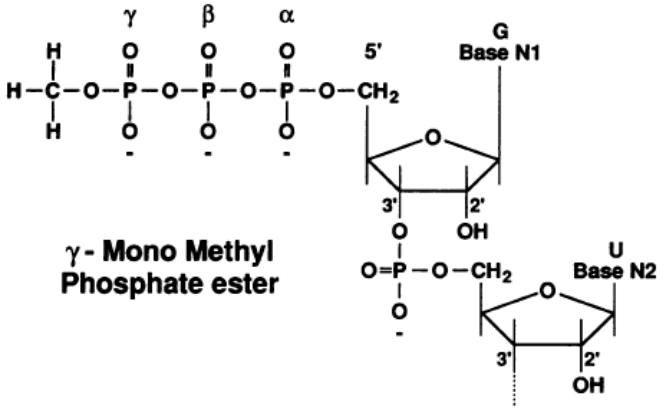


Figure 1.6: Cap structure of some snoRNAs. SnoRNAs, like U3 and U6, produced by RNA polymerase III contain a 5' γ-monomethyl phosphate cap. Figure is taken from (Singh & Reddy, 1989).

1.3.5 Proteins associated with snoRNAs

Both box C/D and box H/ACA snoRNAs associate with four conserved core proteins to form functional box C/D or box H/ACA small snoRNPs. The assembly of the box C/D snoRNP is a hierarchical process and can start as soon as the C/D motif is formed. The C/D motif recruits the 15.5-kDa protein/Snu13p which binds to the internal loop (K-turn) and potentially changes the conformation of the snoRNA allowing the remaining snoRNP proteins Nop56, Nop58 and the Nop1p/fibrillarin methylase to bind (Figure 1.4) (Watkins *et al.*, 2002). Interestingly, only Nop56 and Nop1p/fribrillarin are able to bind to the C'/D' motif which results in structurally asymmetrical snoRNPs (Cahill *et al.*, 2002). This asymmetry might be explained by the different functions of the C/D and the C'/D' motifs. While the C'/D' motif merely guides the methylation of its target sequence, the C/D motif is additionally responsible for the stability and nucleolar localization of the snoRNP (Makarova & Kramerov, 2007). In Archaea, the C'/D' motif in most snoRNAs is highly conserved and this might be the reason for the binding of two complete snoRNP protein sets (Henras *et al.*, 2004b).

Gar1, Nhp2, Nop10 and Dyskerin/Cbf5p/Nap57 are the four proteins that associate and directly interact with the box H/ACA snoRNAs (Figure 1.3B). Unlike the hierarchical assembly of box C/D snoRNPs, box H/ACA snoRNPs form without a step-

by-step procedure because the proteins do not need the RNA to interact with each other (Henras *et al.*, 2004a). Whereas Gar1 stabilizes the snoRNA-target RNA interactions, Dyskerin/Cbf5p/Nap57 is most likely the pseudouridine synthase. The function of the two remaining proteins is not really understood but they are essential for the function of snoRNPs (Lafontaine & Tollervey, 1998; Meier, 2005). In addition to the core snoRNP proteins, around 10 accessory proteins are known to be involved in both assembly and transport of snoRNPs (Meier, 2005).

1.3.6 Organization of snoRNA genes

Methylation and pseudouridylation of many different RNAs is crucial for their biosynthesis. The main target for modification is rRNA where in higher eukaryotes around 100 are 2'-O-ribose methylated and another 100 are pseudouridylated. For each site a specific snoRNA containing the matching antisense element is necessary. It is, therefore, not surprising that more than 100 different snoRNAs have been identified with new ones being frequently discovered. Different box H/ACA snoRNAs were particularly difficult to identify due to their very short conserved sequences, but advances in bioinformatics and genomics have greatly aided their identification. The number of different snoRNAs varies among different species. For instance, *Saccharomyces cerevisiae* has about 81 different snoRNAs (Torchet *et al.*, 2005), while approximately 150 are known in *Arabidopsis thaliana* (Brown *et al.*, 2003a), 118 have been discovered in *Chlamydomonas reinhardtii* (Chen *et al.*, 2008), 217 have been detected in the genome of platypus (Schmitz *et al.*, 2008), and 119 have been found so far within the genome of *Drosophila melanogaster* (Huang *et al.*, 2005). Furthermore, some snoRNAs are transcribed by more than one gene. For example, *Chladomonas reinhardtii,* contains 322 snoRNA genes which encode 118 snoRNAs that modify 158 target sites (Chen *et al.*, 2008). These genes are spread throughout the genome and their organisation varies greatly between different eukaryotes (Figure 1.7).

SnoRNA genes can be found as single genes or in polycistronic clusters. They can be located between protein-coding genes (intergenic) and transcribed independently having their own promoter, or they can be located within introns of protein-coding genes,

relying on the transcription of their host-gene (Figure 1.7). While in animals and yeast the majority of snoRNAs are transcribed from single genes, polycistronic clusters are predominant in plants (Brown *et al.*, 2001; Chen *et al.*, 2003; Chen *et al.*, 2008). Furthermore, only plants and, surprisingly, *Drosophila* contain intronic polycistronic clusters. In vertebrates, most snoRNA genes are found in introns and only some are intergenic. However, the opposite seems true in yeast. Only seven intronic genes and five gene clusters containing 17 genes have been found in the intron-poor *Saccharomyces cerevisiae* genome and most genes in this species are intergenic (Lowe & Eddy, 1999; Qu *et al.*, 1999).

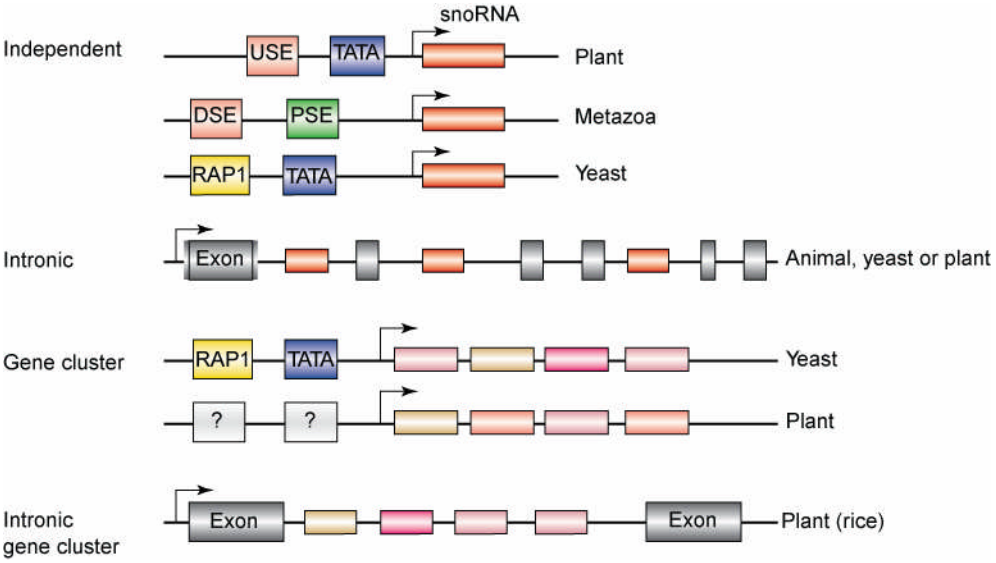


Figure 1.7: Genomic organization and expression of snoRNA genes. SnoRNA genes in plants can be transcribed independently or in introns and can be found as single entities or as a gene cluster (polycistronic). The majority of plant snoRNA genes are organized in clusters, but there is a difference in the percentage of snoRNA genes that are independently transcribed. While the majority of snoRNA genes in animals are located within introns only a few genes in yeast are intronic. Note that intronic gene clusters can also be found in *Drosophila*. Transcription signals and exons are indicated. DSE = distal sequence element; PSE = proximal sequence element; USE = upstream sequence element. Figure is taken from (Brown *et al.*, 2003a).

For organisms with intronic snoRNAs, it appears that generally never more than one snoRNA gene is present within an intron (one-snoRNA-per-intron rule) (Maxwell & Fournier, 1995; Liang *et al.*, 2002). An exception to this rule is found in *Drosophila* which contains a high number of intronic genes and also 17 intronic gene clusters. In this case, only box H/ACA snoRNA genes are organised in clusters whereas box C/D snoRNA genes strictly follow the one-snoRNA-per-intron organization (Huang *et al.*, 2005). In plants, the majority of snoRNA genes (> 80 % in *Arabidopsis* and rice) are found in polycistronic clusters, but the number of intronic gene clusters varies between different species. For instance, whereas the majority of genes/gene clusters in *Arabidopsis* have their own promoter, only about 20 genes, either single or clustered are found in introns (Brown *et al.*, 2008). In contrast, about half of the gene clusters in rice and about 90 % of the gene clusters in *Chlamydomonas* occur in introns of protein-coding genes (Brown *et al.*, 2003a; Brown *et al.*, 2008; Chen *et al.*, 2008). The low number of intronic genes/gene clusters in *Arabidopsis* might be explained by the small average intron size of about 170 nt. In comparison, the average sizes of introns in rice and *Chlamydomonas* are about 360 and 373 nt, respectively (Yu *et al.*, 2002; Merchant *et al.*, 2007; Chen *et al.*, 2008). Thus, small introns appear unsuitable for accommodating a snoRNA gene/gene cluster. In *Drosophila*, for instance, box C/D snoRNA genes can be found only in introns longer than 150 nt (Huang *et al.*, 2005).

SnoRNA gene clusters usually contain two to seven genes, although one gene cluster in rice contains around 42 genes (Chen *et al.*, 2003). Some gene clusters contain copies of the same gene and are referred to as homologous gene clusters. Other clusters contain different genes (heterologous gene clusters) and there are examples of clusters with a combination of both (examples of these clusters are shown in Figure 1.11). For instance, two-thirds of the snoRNA gene clusters found in *Arabidopsis* and rice contain different genes. In contrast, more than 77 % of snoRNA gene clusters in *Chlamydomonas reinhardtii* consist of homologous clusters, which indicate extensive local tandem duplications (Chen *et al.*, 2003; Chase *et al.*, 2005; Chen *et al.*, 2008).

1.3.7 Evolution of snoRNA genes

RNAs which guide RNA modification are found in eukaryotes and archaea, but not in bacteria, suggesting that this ancient site selection mechanism originated in the common ancestor of eukarya and archaea. In bacteria, the few RNA modifications (four methylations and 10 pseudouridylations in rRNA) are carried out using site-specific enzymes (Lafontaine & Tollervey, 1998; Gaspin *et al.*, 2000; Ofengand *et al.*, 2001; Bachellerie *et al.*, 2002; Omer *et al.*, 2003; Leppik *et al.*, 2007; Ero *et al.*, 2008). In archaea and eukaryotes where there is a need to modify more sites in rRNA, perhaps to fine-tune the efficiency of the ribosome, a different system for producing site-specific modifying enzymes evolved. This system, the snoRNPs, could use the same set of proteins for the modification of every target site just by changing the site recognition element. These recognition elements are now known as antisense elements in snoRNAs and sRNAs, respectively. This ancient mechanism appears to be successful as it has not changed a great deal as can be seen in the high conservation between archaea and eukarya guide sequences (Dennis *et al.*, 2001; Omer *et al.*, 2003; Dennis & Omer, 2005). To guide the tens or hundreds of modifications requires large sets of different snoRNAs/sRNAs which have evolved by duplications, mutations and selection of snoRNA/sRNA genes.

Generally, there are two modes of gene/gene cluster duplication: genes/gene clusters can be duplicated close to their origin (cis-duplication) or to a distant location, either on the same or a different chromosome (trans-duplication) (Figure 1.8). For instance, in vertebrates where the majority of snoRNA genes are single and intronic, cis-duplication has occurred when copies of the same intronic gene are found in different introns, neighboring ones or ones that are further away. Trans-duplication is detected when the same snoRNA gene is found within an intron of a different gene. Both duplication modes, although cis-duplication is far more frequent, might act on the same gene causing many widespread copies to be produced. For instance, in *Platypus*, paralogues of a box C/D snoRNA gene are found within the ribosomal protein S13 (RS13) gene as well as in a heat-shock protein 8 (Hsp8) gene and it has been concluded that a single trans-duplication occurred followed by further cis-duplication within one of the two genes (Figure 1.8) (Schmitz *et al.*, 2008). It should be noted that it is crucial for

independent genes/gene clusters that their transcription elements are included in trans-duplication except when they are copied into existing gene clusters or introns. An example of trans-duplication followed by extensive cis-duplication (tandem repeats; see below) is seen in the HBII-52 cluster in humans. HBII-52 is a box C/D snoRNA which might regulate the alternative splicing of the serotonin receptor in human brains. It is thought that a snoRNA gene evolved in an intron of the SNRPB gene which was duplicated and gave rise to SNRPB including the snoRNA now called SNORD. SNORD is also a box C/D snoRNA which might guide methylation of the 28 rRNA. The snoRNA within the SNRPN, however, appears to be extensively duplicated resulting in 42 nearly identical copies (Yang *et al.*, 2006; Nahkuri *et al.*, 2008). In plants, where the majority of snoRNA genes occur in gene clusters, duplications are quite complex. Genes within a cluster, parts of a cluster or the whole cluster can be duplicated. Genes or gene clusters cis-duplicated adjacent to each other are called tandem repeats and are quite common and might be responsible for the origin of gene clusters (Brown *et al.*, 2001; Brown *et al.*, 2003a). A quite impressive case of tandem repeats is known from rice where a cluster (cluster 17) contains 42 genes which have arisen by five tandem repeats (Chen *et al.*, 2003).

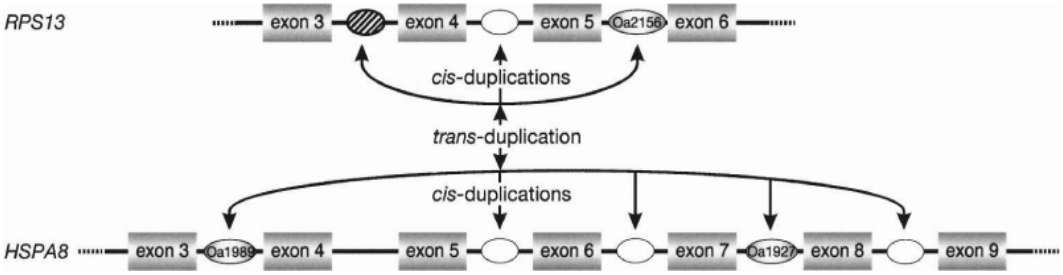


Figure 1.8: Cis- and transduplication of platypus box C/D box snoRNA paralogues. Only one transduplication occurred followed by cis-duplication in one of the two genes. Filled ovals = platypus cDNA library snoRNAs; open ovals = snoRNAs found by blast search; hatched oval = non-functional in platypus but functional in human, mouse and cow. Figure is taken from (Schmitz *et al.*, 2008).

At least 50% of the snoRNA genes in plants have two to four copies produced by the duplication modes described above. Furthermore, polyploidy is prevalent in plants and increases the allelic variants resulting in potential snoRNA redundancy. All in all, there are many genes and alleles producing snoRNAs which target the same nucleotide providing increased chances to accumulate mutations and, thus, generate new antisense sequences leading to new target sites which might be established under selection. For instance, two *Arabidopsis* snoR20 variants, located on two different chromosomes, target neighbouring sites of rRNA (Figure 1.9A). Mutations of an antisense element could also lead to the modification of a site that is distant from the original one as shown by the snoR16 variants (Figure 1.9B). These variants have two antisense elements, the one adjacent to the D' box differs from the other variant by a few nucleotides and targets a completely different site (25S:Um2445 and 25S:Um36, respectively) (Brown *et al.*, 2001; Qu *et al.*, 2001; Brown *et al.*, 2003a). Thus, two or more snoRNA variants might have diverged sufficiently (accumulation of nucleotide changes and indels) to become distinct snoRNAs. An example of the generation of novel snoRNA genes was provided by Brown et al. (2001). In *Arabidopsis* the double guide snoR15 gene was tandemly duplicated after transduplication to another chromosome. Due to mutation one snoR15 variant lost the function of the antisense element adjacent to the D box becoming the snoRNA gene U16, while the other variant lost the box D' antisense element and became the snoRNA U55 gene (Figure 1.9C) (Brown *et al.*, 2001; Brown *et al.*, 2003a). It should be noted that mutations and selection could result in the loss of snoRNA function as well as leading to the production of non-functional pseudogenes which might be lost in time.

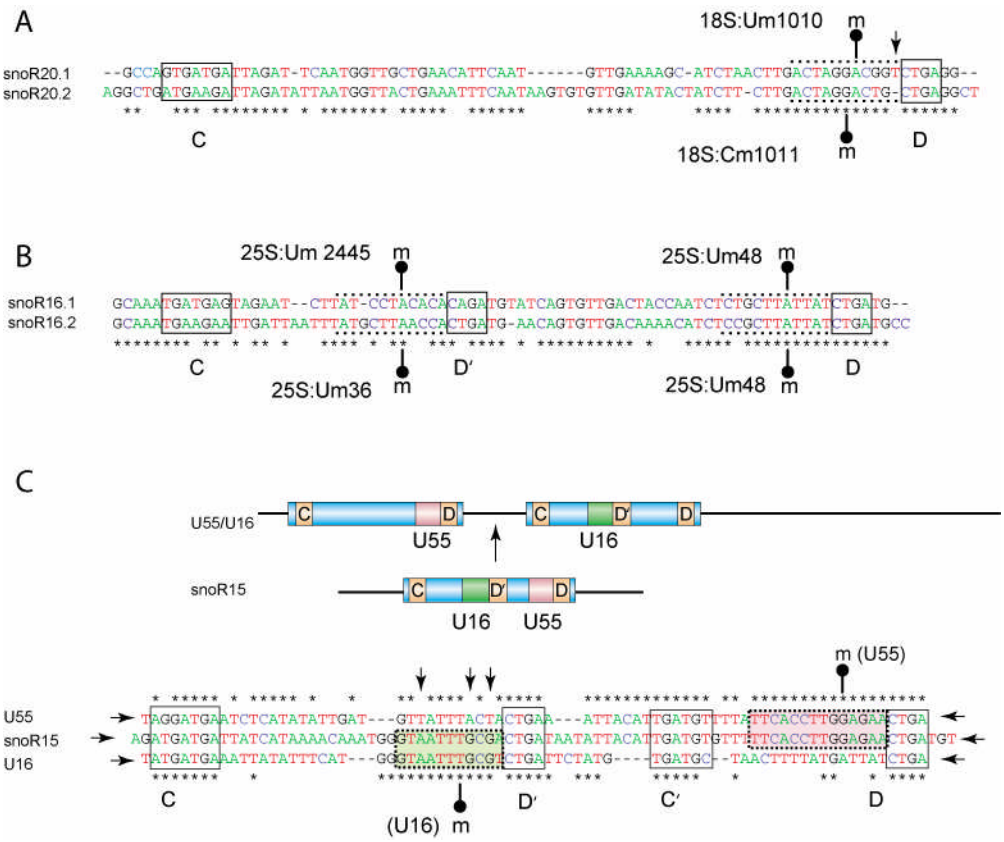


Figure 1.9: Evolution of new snoRNA target sites and genes. A: the two snoR20 variants target the same rRNA but modify neighbouring sites. B: the two antisense elements adjacent to the D' box of the two snoR16 variants differ in some nucleotides and base-pair with different rRNA sites. C: Transduplication followed by tandem duplication and loss of different antisense elements gave rise to U55 and U16 snoRNA genes (Barneche *et al.*, 2001; Qu *et al.*, 2001). m = methylation site. Figure is taken from (Brown *et al.*, 2003a).

Other mechanisms to alter snoRNA gene diversity are gene conversion and unequal crossing over, which mostly lead to the loss of snoRNA genes and reorganisation of snoRNA gene clusters (Barneche *et al.*, 2001; Brown *et al.*, 2001; Qu *et al.*, 2001). It is also feasible that due to unequal crossing over two different adjacent genes might fuse leading to a new chimeric snoRNA, as has be shown for ITS (see Alvarez & Wendel, 2003).

Comparisons of snoRNAs in different plant species have shown variation in the degree of conservation at the level of gene sequence and gene cluster organisation. For example, many orthologous genes show a high degree of conservation in their guide and box sequences while some paralogues appear to exhibit a high degree of variation. A high level of conservation indicates the action of purifying selection on the function of modifying specific target sites (Schmitz *et al.*, 2008), whereas high sequence variation of paralogues might reflect active divergent evolution of snoRNAs. Not only are single snoRNA genes conserved between species, but the gene-order of some gene clusters is conserved as well. Furthermore, some gene clusters contain the same genes and sometimes other genes, which may be positioned in different orders in mixed clusters. In other species, however, these genes can be dispersed across the whole genome (dispersed cluster). For instance, the snoRNA gene cluster 7 in rice contains the same snoRNAs genes and gene order as does snoRNA gene cluster 21 in *Arabidopsis*. However, while all of the cluster 7 genes can be found in rice intronic clusters 47 and 48, these clusters also contain another gene, snoR159, while cluster 47 also contains a copy of U18. These genes are located in *Arabidopsis* cluster 39 together with an inserted U54 gene (Figure 1.10). An example of dispersed cluster is shown in Figure 1.11. Some rice clusters (6, 15 and 17) have a 7 snoRNA gene core structure which in *Arabidopsis* is broken up into different clusters (40, 42 and 43) (Figure 1.11). The gene order of different clusters, either within the same or between different species, suggests that they are often subject to rearrangement in their evolution (Chen *et al.*, 2003). An examination of snoRNA gene organisation in different species might provide insights into the reorganisation and transposition processes that occur during the evolution of different plant lineages (Brown *et al.*, 2003a).

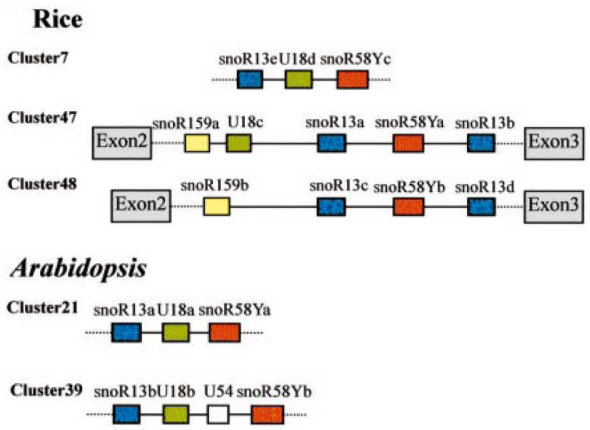


Figure 1.10: Schematic illustration of conserved and rearranged gene order in rice and *Arabidopsis*. Rice cluster 7 appears to be conserved because a similar cluster is found in *Arabidopsis* (cluster 21). All cluster 7 genes are found in rice cluster 47 (mixed cluster) and *Arabidopsis* cluster 39, but either in a different order or with an additional gene inserted (U54). There is no U18 gene present in rice cluster 48. Figure is taken from (Chen *et al.*, 2003).

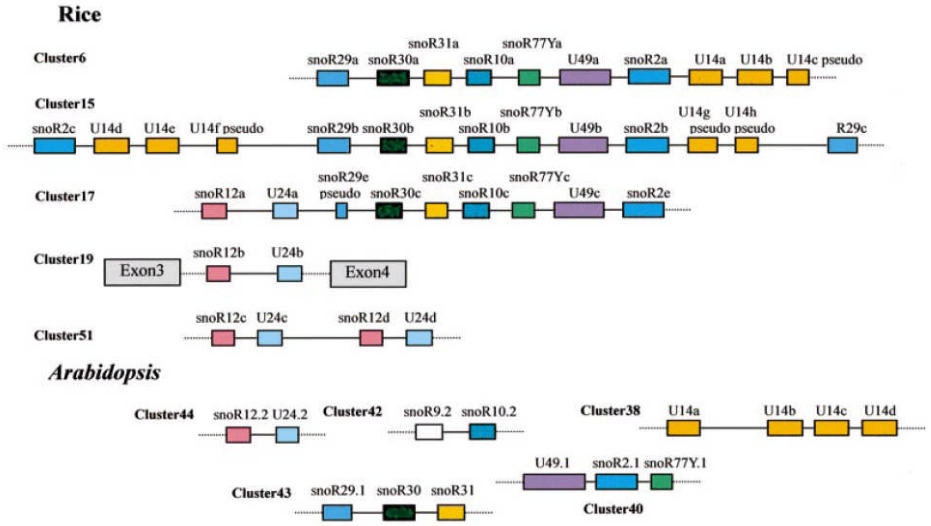


Figure 1.11: Schematic illustration of dispersed gene clusters in *Arabidopsis* relative to rice. Rice clusters 6, 15 and 17 have a 7 snoRNA gene core structure which appears to have broken up into several clusters (40, 42 and 43) in *Arabidopsis*. Some of the clusters found in *Arabidopsis* have the same gene order (e.g. cluster 43 and 44) while others have a mixed (cluster 40) gene order compared to rice. While most of the clusters consist of heterologues (heterocluster), cluster 38 is a good example of a homocluster. Figure is taken from (Chen *et al.*, 2003).

1.3.8 Transcription and processing of snoRNAs

As described above, the organisation of snoRNA genes within the genome differs greatly between various organisms and, thus, several modes of transcription and processing are used (Figure 1.12). Independently transcribed single genes, which are present in animals, yeast and plants contain the snoRNA coding region flanked by their own promotor, enhancer and terminator and, thus, limited processing is necessary (Maxwell & Fournier, 1995; Brown & Shaw, 1998; Brown *et al.*, 2003b). The processing of the majority of single intronic genes in animals depends on splicing which involves the formation of a lariat structure of the snoRNA containing intron. After the linearization of the lariat structure and exonucleolytic trimming (5' and 3' ends) the snoRNP is released (Figure 1.12) (Leader *et al.*, 1997; Filipowicz & Pogacic, 2002; Brown *et al.*, 2008). Most of these snoRNA genes are hosted in introns of protein-coding genes involved in nuclear function and biogenesis of ribosomes. In *Drosophila* and humans, however, some snoRNA genes can be found in introns of genes which do not encode for proteins, but still depend on the intron structure and splicing (Huang *et al.*, 2004; Huang *et al.*, 2005). Processing of independent transcribed polycistronic clusters (in yeast and plants) require a mechanism to separate the several snoRNAs from the snoRNA precursor (Figure 1.12). In yeast, the intergenic spacers contain loop-structures which are cleaved by RNase III. After exonucleolytic trimming of the 5' and 3' ends, the mature snoRNP is generated (Chanfreau *et al.*, 1998; Qu *et al.*, 1999). Plants also contain intronic snoRNAs where some are single genes in an intron, like yeast and mammals, but others are organized into intronic gene clusters. This organization requires a different mode of processing which does not depend on the splicing process and where individual pre-snoRNAs are released by endonucleolytic cleavage. After exonucleolytic trimming (5' and 3' ends) the mature snoRNPs are formed. Individual snoRNAs from intronic polycistronic clusters require endonucleolytic activity but there is no evidence for any formation of stem-loops as in yeast and it is unknown which endonuclease is responsible for the cleavage (Leader *et al.*, 1997, 1999; Brown *et al.*, 2003a; Brown *et al.*, 2008). *Drosophila* also contains intronic box H/ACA snoRNA clusters. The introns are spliced out and the individual snoRNAs are separated and trimmed by endo- and exonuclease, respectively (Figure 1.12). Similar to plant snoRNA spacers, no specific secondary structure was found and, thus, it is quite

likely that endonucleases recognize the structure of the box H/ACA snoRNAs (Huang *et al.*, 2004). The independence of splicing in plants enables them to produce mature snoRNAs and, thus, process various RNAs, even in extreme circumstances when splicing activity is decreased or shut down (Brown *et al.*, 2003a). Similarly, in vertebrates, although the majority of intronic snoRNAs depend on splicing, on rare occasions they are processed independently. For instance, the heat shock protein 70 (hsc70) accommodates U14 snoRNA genes whose processing is splicing independent (Chen *et al.*, 2002). These genes, however, are only expressed under heat-shock showing the important role of splicing independent snoRNA production under extreme conditions.

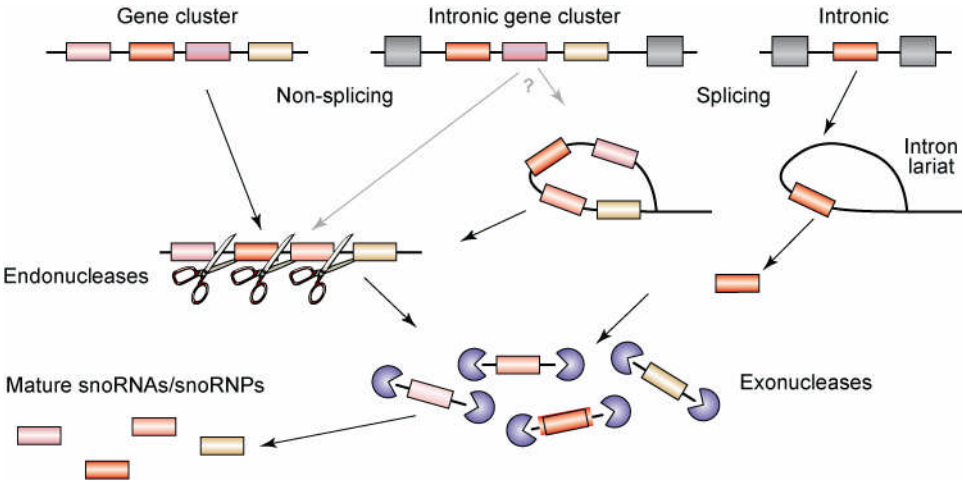


Figure 1.12: Processing of snoRNA genes. An independent gene cluster does not need splicing but endo- and exonucleases for processing. Single intronic genes are processed by splicing, while an intronic gene cluster can be processed by both the splicing and the nonsplicing pathway. After cleavage by endonucleases (nonsplicing), exonucleolytic trimming is necessary. Figure is taken from (Brown *et al.*, 2003a).

An entirely novel mode of gene organization, transcription and processing of snoRNAs has been discovered in *Arabidopsis*, but was later also found in other plants (Figure 1.13) (Kruszka *et al.*, 2003). SnoRNA genes of the snoR43 family were found downstream from tRNA-Gly genes forming dicistronic tRNA-snoRNA (tsnoRNA) gene clusters, which are most likely transcribed from the tRNA promoter by RNA polymerase III. The tsnoRNA-precursor is processed by RNase P and RNase Z, thereby releasing the mature

tRNA and snoRNA. The 3' extension of the snoRNA might be endonucleolytically removed and trimmed by exonuclease. The formation of the box C/D motif and association of the snoRNP proteins with the snoRNA might already take place just after transcription of the pre-tsnoRNA (Figure 1.13) (Kruszka *et al.*, 2003).

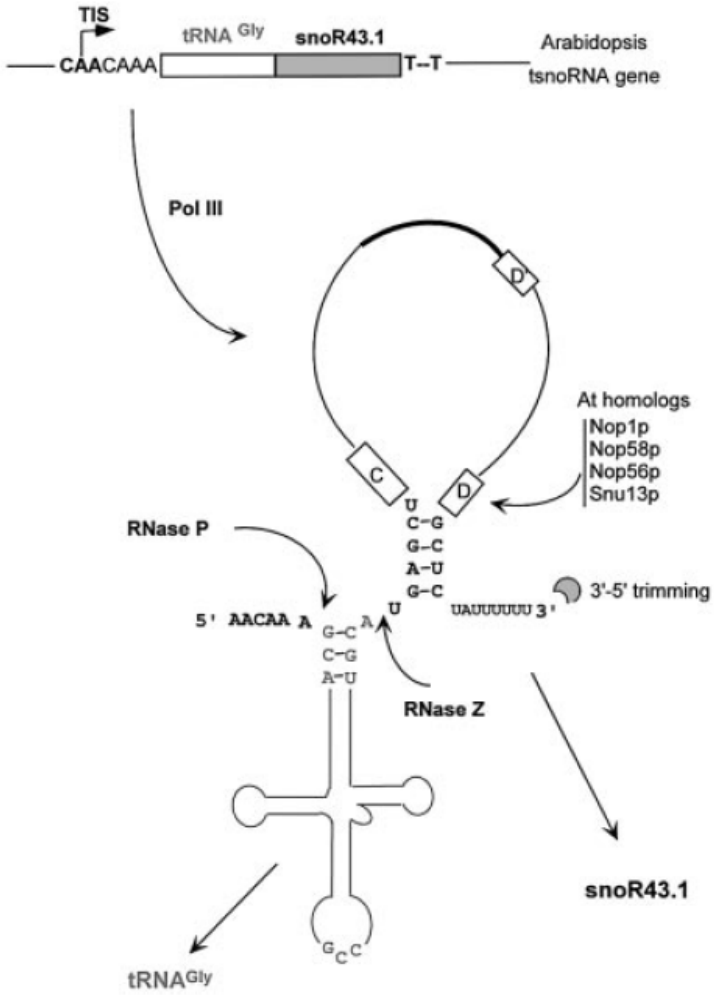


Figure 1.13: Organization, transcription and processing of tRNA-snoRNA dicistronic gene clusters in plants. The snoR43.1 gene is located directly downstream from the tRNA-Gly gene. Using the tRNA promoter, the tsnoRNA precursor is transcribed by RNA polymerase III. The snoRNP core proteins bind to the box C/D motif most likely formed directly after transcription. The 5' extension of the tRNA is removed by RNase P and RNase Z releases mature tRNA and snoRNA by cleavage. The 3' extension of the snoRNA is endo- and exonucleolytically processed. Figure is taken from (Kruszka *et al.*, 2003).

1.3.9 snoRNAs involved in the modification and processing of rRNAs

The majority of snoRNAs are involved in the modification of ribosomal RNAs (rRNA) of the cytoplasmatic ribosomes during their biosynthesis. These modifications (2'-O-ribose methylation and pseudouridylation) occur co-transcriptionally with rRNA transcription and are necessary for the right folding of the rRNA, RNA-RNA and RNA-protein interactions (Kiss *et al.*, 2004; Zemann *et al.*, 2006). Additionally, some snoRNAs (eg U3, U8, U14, U17, U22 and RNAse MRP RNA) play an important role in the processing of the rRNA-precursor to mature 5.8S, 18S and 25/28S rRNAs, which involves endonucleolytic cleavages of the internal and external transcribed spacers (ITS and ETS). A few of these snoRNAs, for example the U3 or the U14 snoRNA are conserved in all eukarya, whereas others such as the U8 snoRNA are only found in vertebrates (Brown & Shaw, 1998; Venema & Tollervey, 1999; Lafontaine & Tollervey, 2001; Brown *et al.*, 2003a; Chen *et al.*, 2008). During processing the rRNA precursor is associated with ribosomal and non-ribosomal proteins as well as with snoRNPs forming 90S pre-ribosomal particles (prp) which will be split into 40S and 60S prps (Henras *et al.*, 2008).

A crucial snoRNA, involved in more than one process during ribosome biogenesis, is U3. As shown in *Xenopus*, U3 snoRNAs are necessary for pre-rRNA cleavages in the 5'ETS, ITS1 and at the 5' end of the 18S region, although no evidence for any endonucleolytic activity was found (Borovjagin & Gerbi, 1999, 2001, 2005). In yeast, U3 associates with the GTPase Bms1p and the putative endonuclease Rcl1p forming a ternary complex and interacts directly with the 90 S prp by base-pairing. GTP bound to Bms1p increases affinity for Rcl1p and Rcl1p bound to GTP-Bms1p raises the affinity for U3 snoRNA. Binding to the 90S prp causes a change in its conformation which might trigger the hydrolysis of GTP to GDP leading to the dissociation of the ternary complex due to the decrease in affinity (Gelperin *et al.*, 2001; Wegierski *et al.*, 2001; Karbstein *et al.*, 2005; Karbstein & Doudna, 2006). The U3 RNA interacts with the 5'ETS pre-rRNA via two distinct hinge regions (Borovjagin & Gerbi, 2000). Additionally, U3 snoRNAs have two conserved boxes, A and A', respectively. These two boxes might play an important role in the correct folding of the center core (pseudo-knot) of the 18S rRNA by preventing the wrong base-pairing of two sequence elements which are more than 1 kb apart. Furthermore, by binding to the upstream and the more

downstream elements brings the two base-pairing partners in close spatial proximity leading to the formation of the pseudo-knot (Hughes, 1996; Mereau *et al.*, 1997; Henras *et al.*, 2008).

U14, U17 and U22 are involved in the early cleavage stages of 18S pre-rRNA (Li *et al.*, 1990; Tycowski *et al.*, 1994; Atzorn *et al.*, 2004). For instance, U14 and U17 interact directly with the 35S pre-RNA. U14 box C/D snoRNAs contain the so-called domain A which is necessary for the 18S rRNA procution (Jarmolowski *et al.*, 1990; Liang & Fournier, 1995; Peculis, 1995). There are two putative pre-rRNA binding sequences in U17 box H/ACA snoRNAs, but their interaction with the 35S region is still unknown (Atzorn *et al.*, 2004). Only the U14 snoRNA is involved in both processing and modification of pre-rRNA (Dunbar & Baserga, 1998).

U8 box C/D snoRNAs, only found in vertebrates, are crucial for the processing of the large subunit rRNAs 5.8S and 28S. At the top of the third stem, U8 snoRNAs contain a conserved octamer sequence for binding LSm proteins in all probability inducing modulations in the RNA structure required for stability in both U8 and ribosomal RNA. The conserved sequence at the U8 5’ end binds to the 5’ end of the 28S pre-rRNA which leads to the correct folding of the proximal stem of ITS2 which will be removed by cleavage later on, and an interaction between 5.8S and 28S rRNA (Peculis & Steitz, 1993; Tomasevic & Peculis, 1999; Peculis *et al.*, 2001; Tomasevic & Peculis, 2002; Ghosh *et al.*, 2004). Additional putative U8 snoRNA binding sites in this pre-rRNA region were found by comparative analysis (Michot *et al.*, 1999). A protein binding to U8 snoRNA with high affinity is X29. X29 is a nuclear decapping enzyme and might regulate the level of nuclear RNAs with methylated caps, including snoRNA U8 (Ghosh *et al.*, 2004).

1.4 Investigated snoRNA gene clusters

Thirty eight different snoRNA genes found in 14 gene clusters in the *Arabidopsis thaliana* genome were investigated in this study (Appendix 1, Figure A.1). Most of these gene clusters and also parts of them are present in more than one copy in *A. thaliana*, either on the same and/or different chromosomes, but some clusters are found only once. For instance, while cluster C is present in three copies, two found on chromosome 1 and one on chromosome 4, there is only one cluster K which is found on chromosome 1. Furthermore, the cluster found on chromosome 5 is also present on chromosome 3, but lacks gene U54. Additionally, cluster D can be found three times within the *A. thaliana* genome but the copy on chromosome 4 lacks a large part of the snoR77Ygene.

1.4.1 Composition and location of the gene clusters

The majority of snoRNA gene clusters in *A. thaliana* consist of two to five heterologous genes. Some gene clusters, such as cluster M consist not only of heterologous genes but also contain homologues (Appendix 1, Figure A.1). Most gene clusters contain predominantly box C/D snoRNA genes. However, two gene clusters, K and L, consist of more box H/ACA genes than box C/D ones. The difference in the number of box C/D and H/ACA genes found in the gene clusters investigated is partly due to identification difficulties of the latter (Brown *et al.*, 2003a) which has since been addressed using various approaches (e.g. Huang *et al.*, 2007; Chen *et al.*, 2008; Hertel *et al.*, 2008). Ten gene clusters (A-J) containing 28 box C/D and 2 box H/ACA genes were already identified and characterized in 2001 (Brown *et al.*), whereas four gene clusters (K-N) consisting of 6 box C/D and 8 box H/ACA genes were discovered more recently (Brown et al., unpublished data).

The 14 gene clusters and their putative copies are spread across the whole genome: chromosome 1 contains six clusters, chromosome 2 four, chromosome 3 five, chromosome 4 six, and chromosome 5 contains three snoRNA gene clusters (Appendix 1, Figure A.1).

1.5 Aims of research

The major aim of the research reported in this thesis was to investigate snoRNA genes and gene clusters for their potential application in phylogenetic studies and DNA barcoding. Various snoRNA genes and gene clusters identified in *Arabidopsis thaliana* were compared with a large number of expressed sequence tag (EST) databases from multiple plant species and homologous sequences were aligned and used for designing universal primers (Chapter 3). The universal primers were then tested for amplification in various *Senecio* species and the variation of their amplified products among and within species was assessed by analysing fragment lengths profiles (Chapter 4) and sequence data (Chapter 6). In more detail, the fragment length profiles obtained from various genes and gene clusters were tested for their ability (i) to separate closely and more distantly related *Senecio* species and (ii) to detect hybrids (Chapter 4). As the fragments obtained cannot be clearly assigned to certain orthologous regions the profiles were scored as dominant markers. Sequence data were used to identify orthologues and putative paralogues and to isolate single copy regions which could then be scored as more informative co-dominant markers (Chapter 6). Furthermore, the sequences obtained were also examined for their ability to discriminate closely related species and thus their potential for DNA barcoding.

Another goal of the research was to characterize snoRNA genes and gene clusters in *Senecio* (Chapter 5). The fragment lengths obtained using universal primers were compared with the expected fragment lengths from other species, especially *Arabidopsis thaliana*, and possible snoRNA gene clusters were reconstructed. Sequence data were then used to identify putative gene/gene cluster copies which were not discovered by fragment analysis (Chapter 6).

Many snoRNA genes and gene clusters are present in more than one copy and therefore might be an ideal system for studying gene evolution. Thus, a third goal was to investigate the evolution of certain snoRNA genes/gene cluster using sequence data, especially by comparing putative paralogous sequences. Additionally, the potential of snoRNA gene/gene cluster for examining gene evolution was assessed.

1.6 Study species of Senecio

The Senecioneae is the largest tribe within the Asteraceae family and consists of about 150 genera and more than 3000 species (Nordenstam, 2003, 2007). About 1250 species belong to the genus *Senecio* (Coleman *et al.*, 2003). It is one of the largest of angiosperm plant genera and is found throughout the world apart from Antarctica (Pelser *et al.*, 2007). A large number of *Senecio* species are annual or short-lived perennials. Most have radiate flower heads (capitula) containing a central disc composed of disc florets bordered by an outer whorl of ray florets, e.g. *S. squalidus* L., while a minority produce non-radiate or discoid flower heads that lack ray florets and contain only disc-florets; e.g. *S. vulgaris* L. var. *vulgaris* L. A recently constructed phylogeny of the Sencioneae based on nuclear rDNA ITS sequences has shown that most species currently assigned to *Senecio* form a well supported clade (Pelser *et al.*, 2007). However, the analysis makes clear that a revision of the genus is required involving the addition of some species to the genus and the removal of others.

Most of the work reported in this thesis centred on the species *Senecio squalidus*, *S. aethnensis* Jan. ex DC, *S. chrysanthemifolius* Poiret and *S. cambrensis* Rosser, and *S. vulgaris var. vulgaris*. Other species of *Senecio* were examined where necessary to investigate (i) the variability of the snoRNA gene marker system between closely and distantly related species, and (ii) hybridisation events.

1.6.1 Phylogenetic relationships between test species

To obtain a picture of phylogenetic relationships among species examined in this thesis a neighbour joining tree was generated from internal transcribed spacer (ITS) sequences downloaded from GeneBank (http://www.ncbi.nlm.nih.gov/) for all species, except *S. vulgaris* L. var. *hibernicus* Syme, *S. massaicus* (Maire) Maire and S. *teneriffae* Sch. Bip. (Figure 1.14). The phylogeny identifies a well-supported clade containing the African taxa *S. flavus* (Dcne.) Schultz Bip and *S. engleranus* O. Hoffm., which is sister to a clade containing all other taxa. In the second clade, *S. madagascariensis* Poir. and *S. inaequidens* DC, which are native to South Africa, form a 100 % bs supported clade that

is sister to a 79 % bs supported clade containing two well supported sub-clades (bs = 100 % and 99 %, respectively). While one of the sub-clades consists of *S. vulgaris* var. *vulgaris* and *S. vulgaris* L. ssp. *denticulatus* (O. F. Muell.) P. D. Sell, a winter annual with an Atlantic-Mediterranean-montane distribution (Kadereit, 1984), the other sub-clade contains the remaining species examined, i.e. *S. glaucus* L. ssp *coronopifolius* (Desf.) Alexander, *S. squalidus* ssp. *araneosus* (Emb. & Maire) Alexander, *S. squalidus* ssp. *squalidus*, *S. chrysanthemifolius*, *S. aethnensis*, *S. rodriguezzii* Willk. ex Rodrigo, *S. glaucus* ssp. *glaucus*, *S. mohavensis* A. Gray ssp. *breviflorus* (Kadereit) M. Coleman and *S. mohavenesis* ssp. *mohavensis* (Figure 1.14). Thus, the species used in this study belong to four well supported and distantly related clades: the *S. flavus*-*S. engleranus* clade, the *S. inaequidens*-*S. madagascariensis* clade, the *S. vulgaris* clade and the remaining clade referred as the *S. squalidus* clade (equivalent to clade A in Pelser *et al.* (2007)).

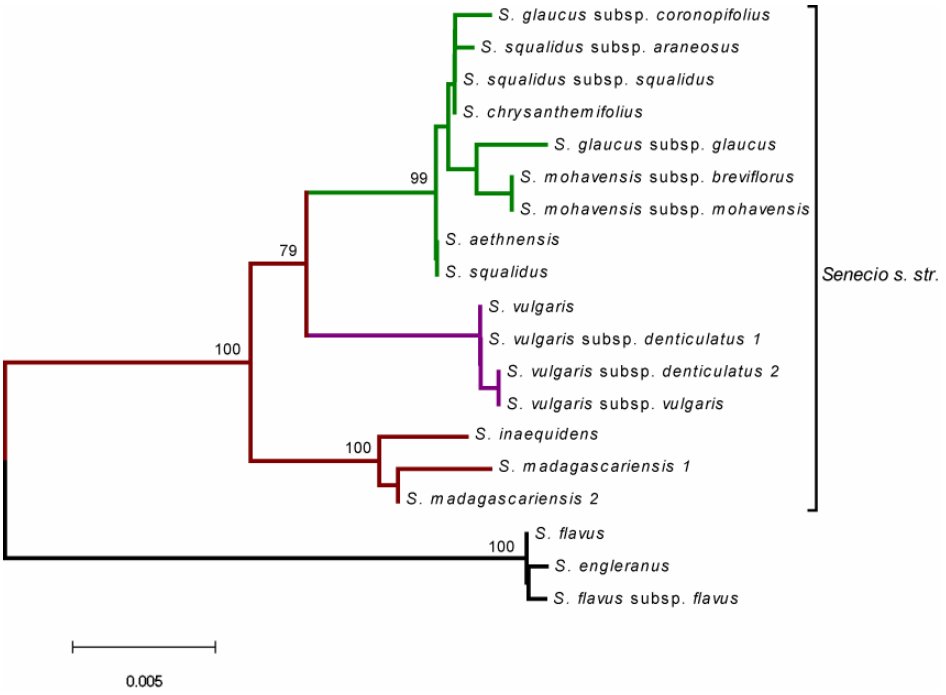


Figure 1.14: Evolutionary relationships of *Senecio* species used in this study. Relationships are based on NJ analysis (Saitou & Nei, 1987) of ITS sequence variation using the Maximum Composite Likelihood method (Tamura *et al.*, 2004). Numbers above branches indicate bootstrap values for the major clades from 1000 replicates (Felsenstein, 1985). Phylogenetic analyses were conducted in MEGA4.1 (Tamura *et al.*, 2007). Sequences are identified in Appendix (Table A.1).

1.6.2 Hybridisation, introgression and polyploidisation

Seven natural *Senecio* hybrids have been recorded in the British flora suggesting that hybridization, introgression and polyploidization may be common in the genus. Four of these hybrids are recognised as new hybrid taxa. One of these, *Senecio squalidus,* is a recently originated homoploid hybrid species derived from plants collected from a hybrid zone between two diploid species *S. chrysanthemifolius* Poiret and *S. aethnensis* Jan. ex DC, (2n = 20), on Mount Etna, Sicily (James & Abbott, 2005) (Figure 1.15). Material from this hybrid zone was introduced to Britain at the beginning of the 18th century (Harris, 2002) and after a period of cultivation and stabilisation in the Oxford Botanic Garden, began to spread via the railway network in the late nineteenth century. The new hybrid species is now well established throughout a large part of Britain (James & Abbott, 2005; Abbott *et al.*, 2009) and is genetically divergent from all hybrids on Mount Etna (James & Abbott, 2005).

A second hybrid taxon, the tetraploid (2n = 40) inland radiate form *S. vulgaris* (var. *hibernicus*), originated via introgression between *S. squalidus* and the tetraploid (2n = 40) discoid form of *S. vulgaris* (var. *vulgaris*) (Ingram *et al.*, 1980; Abbott *et al.*, 1992; Kim *et al.*, 2008) (Figure 1.15). Trow (1912) showed that a single genetic locus controls the presence or absence of ray florets in *S. vulgaris* capitula and there is now full-proof evidence that the ray 'allele' responsible for producing ray flowers was introgressed from *S. squalidus* into *S. vulgaris* var. *vulgaris* leading to the origin of *Senecio vulgaris* var. *hibernicus* (Abbott *et al.*, 1992; Kim *et al.*, 2008).

The third hybrid taxon, the allohexaploid (2n=60) *S. cambrensis,* is known to have originated independently in north Wales and Edinburgh after hybridization between *S. vulgaris* var. *vulgaris* (acting as the maternal parent) and *S. squalidus* (Abbott *et al.*, 1992; Harris & Ingram, 1992b; Abbott & Lowe, 2004) (Figure 1.15). Hybridization between these two species also gave rise to the fourth hybrid taxon, the fertile tetraploid hybrid *S. eboracensis* (2n = 40) (Irwin & Abbott, 1992; Lowe & Abbott, 2000; Abbott & Lowe, 2004) (Figure 1.15). While *S. eboracensis* has not been recorded in the wild since 2000, *S. cambrensis* is still found in north Wales, although in declining numbers (Abbott *et al.*, 2007; Abbott *et al.*, 2009).

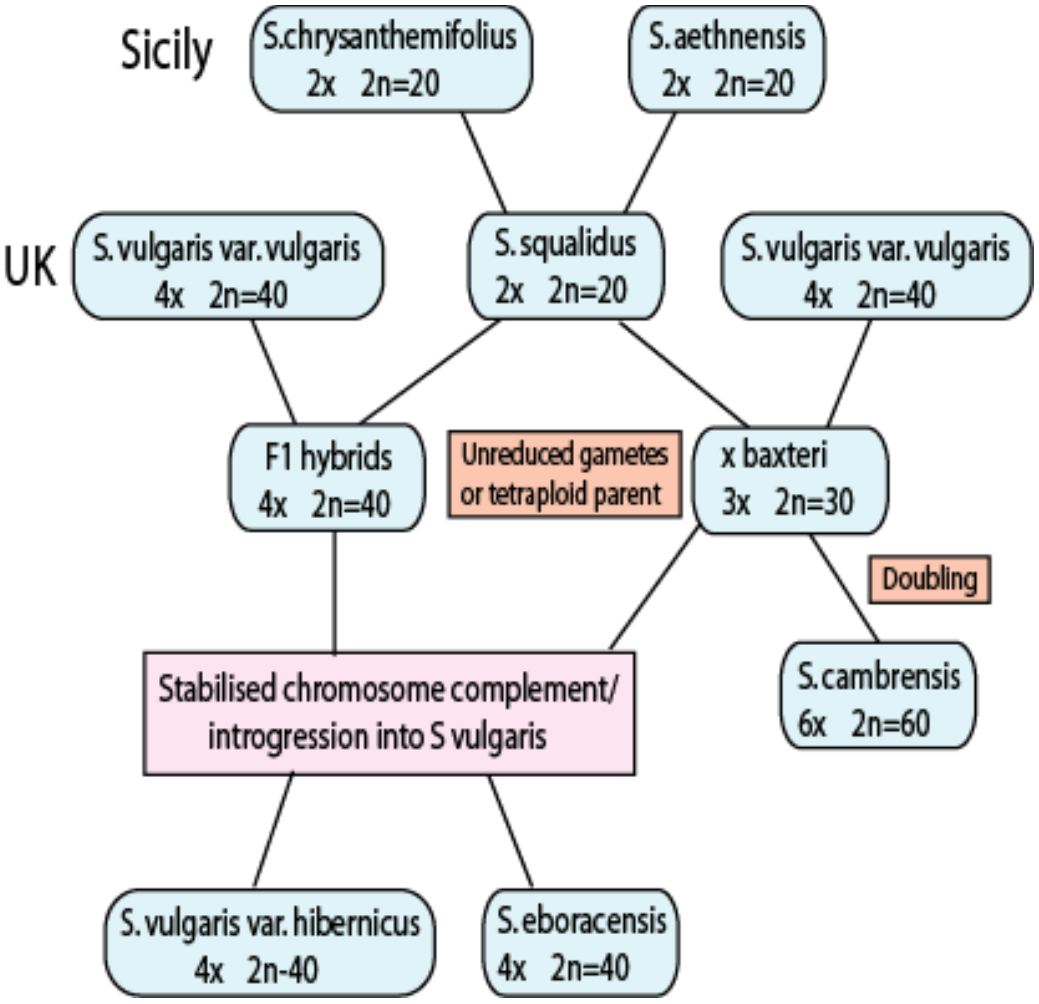


Figure 1.15: Relationships and origins of some British *Senecio* species used in the analyses of the SnoRNA marker system. Hybridisation and introgression was involved in the origin of most of these species. Chromosome numbers are shown below each species name. *S. eboracensis* was not included in the snoRNA analysis (from Lowe *et al.*, 2004).

In addition to the hybrid *Senecio* species that are known to occur in Britain, several other hybrid species in the genus have been recognised from elsewhere. A closely related species to *S. cambrensis* is the allohexaploid, *S. teneriffae* Schultz Bip. (2n = 60), which is endemic to the Canary Islands. This is believed to have originated from a cross between *S. vulgaris* and the diploid (2n = 20) *S. glaucus* L. (Lowe & Abbott, 1996) and represents a rare example of allopolyploid speciation on an oceanic island (Lowe *et al.*, 2004). *Senecio glaucus* is also believed to have been involved as a parent in the origin of

two other allopolyploid species, the hexaploid *S. hoggariensis* (2n = 60) a native of North Africa, and the tetraploid *S. mohavensis* A. Gray ssp. *breviflorus* (Kadereit) M. Coleman (2n = 40), native to south-west Asia. In both instances, the other parent is believed to be *S. flavus* (Dcne.) Schultz Bip. (Comes & Abbott, 2001; Coleman & Abbott, 2003; Coleman *et al.*, 2003; Kadereit *et al.*, 2006). There is good molecular evidence that *S. hoggariensis* consists of two diploid genomes of *S. glaucus* and one diploid genome of *S. flavus*, with *S. flavus* acting as the female parent. In the case of *S. mohavensis* ssp. *breviflorus, S. glaucus* is thought to be the female parent (Kadereit *et al.*, 2006). A long distance dispersal event from south-west Asia to North America is believed to have enabled *S. mohavensis* to colonize western North America and led to the origin of ssp. *mohavensis* (Coleman *et al.*, 2003).

1.6.3 Invasive *Senecio* species

Some *Senecio* species, such as *S. squalidus, S. inaequidens* DC, *S. pterophorus* and *S. madagascariensis* Poir., are highly invasive and widespread. *Senecio inaequidens, S. madagascariensis* and *S. pterophorus* are native to South Africa, but have become highly invasive weeds mainly in parts of the Northern Hemisphere, Southern Hemisphere and Australia, respectively. The taxonomy of *S. inaequidens* and *S. madagascariensis* is unclear and it is feasible that the two species have undergone introgressive hybridization in South Africa. It is important to be able to identify these two species and also their putative hybrids for conservation and control purposes (Le Roux *et al.*, 2006).

Chapter 2: Material and Methods

2.1 Plant material

Seed collected from wild plants by others was available for all species examined in the present study and was held in stock in the Labaratory of the School of Enviromental and Evolutionary Biology, University of St. Andrews.

Plants were grown in a green house from either stored seeds or seeds obtained by artificial crossing. DNA was extracted from leaves of 154 different accessions (ac) comprising 16 species/subspecies/varieties, *S. aethnensis* x *S. chrysanthemifolius* hybrids sampled from various populations across the hybrid zone between these two species on Mount Etna, Sicily, and F1 hybrids produced from artificially crossing the same two species (Table 2.1).

Table 2.1: Geographic location, site description, coordinates (latitude and longitude) altitude and number of individuals per population (N) collected from different species used in the present study.

Species	N	Location	Latitude			Longitude			Altitude
			00	00'		00	00'		m
S. aethnensis		Sicily, Mount Etna,							
	2	**Cisternazza**	37	44	N	15	1	E	2600
	5	**Piano Provenzana**	37	47	N	15	1	E	1800-2000
	10	**Piano Provenzana**	37	47	N	15	1	E	2181
	5	**Rifugio Sapienza**	37	42	N	14	59	E	2000
S. chrysanthemifolius		Sicily, Mount Etna,							
	6	**Pedara**	37	37	N	15	4	E	650
	5	**Randazzo**	37	53	N	14	57	E	750
	3	**Catania**	37	32	N	15	5	E	~100
***S. aethnensis* x *S. chrysanthemifolius* hybrid**		Sicily, Mount Etna,							
	2	**Monte Albano**	37	43	N	14	54	E	1425
	2	**Rifugio Sapienza**	37	40	N	14	59	E	1329
	4	**Piano Provenzana**	37	48	N	15	4	E	1603

	8	**Rifugio Sapienza**	37	42	N	14	60	E	1928
	1	**Monte Albano**	37	44	N	14	55	E	1530
	2	**Rifugio Sapienza**	37	41	N	14	59	E	1515
***S. aethnensis* x *S. chrysanthemifolius* F1**	5								
***S. chrysanthemifolius* x *S. aethnensis* F1**	5								
S. squalidus		UK,							
	2	**Edinburgh**	55	58	N	3	7	W	<50
	5	**Edinburgh**, Leith	55	58	N	3	7	W	<50
	5	**Oxford**	51	45	N	1	9	W	<50
	5	**St Helens**	53	23	N	2	45	W	<50
	5	**Cardiff**	51	25	N	3	9	W	<50
	4	**Pentre**	51	39	N	3	29	W	<200
	1	**Summerhill**	53	5	N	3	2	W	<150
	2	**York**	53	53	N	1	4	W	<100
S. vulgaris* var. *hibernicus		UK,							
	1	**New Brighton**	53	11	N	3	7	W	<150
var. *denticultatus*									
	1	**Jersey**	49	13	N	2	8	W	<100
var. *vulgaris*									
	2	**Edinburgh**, Leith, Salamander St	55	58	N	3	7	W	<50
	2	**Edinburgh**, Newhaven	55	59	N	3	12	W	<50
	1	**York**	53	53	N	1	4	W	<100
	2	**Cardiff**	51	25	N	3	9	W	<50
	2	**Pentre**	51	39	N	3	29	W	<150
		Egypt**,**							
	1	**Tanta**	30	48	N	31	0	E	<50
S. cambrensis		UK							
	5	**Edinburgh**, Leith	55	58	N	3	7	W	<50
	1	**Wrexham**	53	2	N	3	0	W	<100
	1	**Mochdre**	53	15	N	3	47	W	<50
	2	**Pentre**	51	39	N	3	29	W	<150
	1	**Ffrith**	53	5	N	3	4	W	<150
	1	**New Brighton**	53	11	N	3	7	W	<150
	1	**Chirk**	52	56	N	3	3	W	<150
S. teneriffae		Tenerife,							
	2	**La Palma**	28	43	N	17	54	W	1350
	1	no location							

S. flavus									
	1	Morocco	31	51	N	7	6	W	
	1	Canary Islands	28	17	N	16	35	W	
	1	Egypt, **Sinai Peninsula**	30	52	N	32	20	E	
S. glaucus* ssp. *coronopifolius									
	1	Morocco, **Onafka**	31	51	N	7	6	W	
	1	Israel, **Khirket Mezin**	31	40	N	35	26	E	<50
	1	Israel, **Mizpe Ramon**	30	37	N	34	48	E	<900
S. engleranus									
	1	no location							
	3	S Africa, Namibia	22	50	S	18	26	E	
S. massaicus									
	1	Morocco, **Sous river**	31	51	N	7	6	W	
	1	Tenerife, **El Medano**	28	17	N	16	38	W	
S. mohavensis **ssp. *breviflorus***									
	1	Israel, **Khirket Mezin**	31	40	N	35	26	E	<50
	1	Israel, **Paran Ha Neshar**	31	48	N	34	46	E	<100
ssp. *mohavensis*									
	1	USA, California, **San Bernadino**	34		N	117		W	
	1	USA, Arizona, **Painted Rock**	34		N	111		W	
S. madagascariensis									
	1	Mad-Kitang							
	1	S. Africa 14.2 22-**East London**	26	28	S	29	6	E	1650
	4	Australia, **Killarney**, Stumkats‘	28	20	S	152	18	E	500
	2	Australia, **Lamington NP**, O’Reillys	28	13	S	153	8	E	850
	2	Australia, SE Queensland, **Springbrook**	28	13	S	153	16	E	750
	2	S Africa, **Niki Nana**	33	1	S	27	55	E	<50
	3	S Africa, **Haga Haga**	32	46	S	28	15	E	<50

Seeds were sown on moist filter paper in transparent plastic boxes kept either on a window sill or in a growth chamber with a 16 hour photoperiod at about 20° C. Young seedlings were transferred to pots containing a 3:1 compost and gravel mixture and grown in the greenhouse. Additional artificial light was supplied and the temperature maintained at about 22° C. Plants were protected from parasites, watered, and treated with fertilizer as necessary.

2.2 DNA extraction

Total DNA was extracted from either frozen, dried or fresh leaves using a modified 2x CTAB (hexadecyltrimethyl ammonium bromide) minipreparation extraction method (Doyle & Doyle, 1987) or alternatively the DNeasy Plant Mini Kit (Qiagen; for procedure see DNeasy Plant Mini Kit protocol p. 24-27; http://www1.qiagen.com/jump/DNeasyKitsReferences.aspx).

2.2.1 2x CTAB procedure

About 200 to 250 mg of fresh or about 40 to 80 mg dried leaf tissue was transferred to a 2 ml reaction tube, flash frozen with liquid nitrogen and pulverized to a fine powder using a plastic pestle. One ml of 2x CTAB extraction buffer (0.1 M Tris pH 8.0, 1.4 M NaCl, 20 mM EDTA (ethylenediaminetetraacetic acid), 2% CTAB) containing 1-2 % of 2-mercaptoethanol (added before used), pre-warmed 55° C, was added, thoroughly mixed, placed in a 65° C water bath for about 30 minutes and cooled for 10 minutes. 700 µl of CI (chloroform-isoamylalcohol; 24:1) were added to the tube, vortexed, centrifuged for 10 minutes at 13000 rpm (*Biofuge pico*, Heraeus Instruments) and then the aqueous supernatant (upper phase) containing the DNA was transferred to a clean 2 ml reaction tube. Another 700 µl of CI were added to the collected supernatant, centrifuged and transferred, as described above. RNA was removed by addition of 3 µl RNase (10 mg/ml), followed by incubation at 37° C for 1 hour. DNA was precipitated by adding 700 µl (2/3 v/v) ice-cold isopropanol, followed by 3x inverting and incubating at -20° C for 30 minutes to overnight (o/n). To pellet the DNA, the sample was centrifuged for 10

minutes at 13000 rpm. The supernatant was poured away and 500 µl ice-cold 70 % ethanol were added to wash the pellet. The sample was thoroughly mixed and centrifuged for 2 minutes at 13000 rpm. The supernatant was carefully poured away, the tubes inverted on a towel and air dried for at least 20 minutes. The DNA was resuspended in 50 µl TE buffer (10 mM Tris, 1 mM EDTA; pH 8.0) and stored at -20° C.

2.2.2 Agarose gel electrophoresis

2.2.2.1 Preparing the gel

1.5 % Agarose gels were used to quantify the DNA content from DNA extracts, and also PCR/cloning amplification products (see below). Depending on the number of samples examined different gel rigs were used. Small (about 50 ml) gels containing 40 wells and medium (about 150 ml) ones containing 100 wells were employed. DNA was stained using ethidiumbromide (3,8-diamino-5-ethyl-6-phenylphenanthridiniumbromide - EtBr).

A small 1.5 % agarose gel was poured by mixing 0.75 g of electrophoresis-grade agarose (BioGene) and 50 ml 0.5x TBE (tris-borate-EDTA) buffer (44.5 mM Tris, 44.5 mM boric acid, 1 mM EDTA; pH 8.0) in an erlmeyer flask. The mixture was heated in a microwave and regularly mixed until the agarose had dissolved. The mixture was then left to cool to about 50° C on a magnetic stirrer, before adding 2.5 µl of EtBr (10 mg/ml) and carefully mixing. The solution was poured into a comb containing plastic gel mould. Potential bubbles were removed using the wide opening of a yellow tip and the gel was left to set for at least 20 minutes. Combs were removed and the gel was immersed in about 60 ml 0.5 x TBE buffer.

DNA samples (3 to 10 µl) were mixed with 2 µl 6x loading dye (0.25% bromophenol blue, 60 mM EDTA, 30% glycerol) and SDW to a final volume of 12 µl and loaded into the wells produced by the combs. One well, usually the first one, was loaded with 5 µl of 100bp DNA Ladder (Promega), which can be used for estimating both the fragment length and the DNA content. For the quantification of DNA extracts, 25 ng and 50 ng of λ DNA (GIBCO BRL) were also loaded on a gel.

A small gel was run at 63 V, a medium one at 90 V, for about 30 to 40 minutes and afterwards photo-documented using a gel image analysis system (Herolab, E.A.S.Y.

Store software or DOC-008XD - UVItec). The gel picture was used to estimate the DNA content by comparing the band intensity of the samples with the DNA ladder bands and, in the case of DNA extracts with the bands of λ DNA. Fragment sizes were estimated by comparing the sample bands with the DNA Ladder.

2.2.3 Quantifying the DNA extracts using photometry:

As the method described in section 2.2.2 provides only a rough estimate of DNA quantity, various samples were also measured spectrophotometrically using NanoDrop® ND-1000 Spectrophotometer. DNA has its absorptions maxima at 260 nm due to the aromatic rings of its bases and this characteristic is used for determining the concentration of nucleide acids. One µl of a 1:10 sample dilution (1 µl sample + 9 µl SDW (sterile distilled water)) was measured at 260/280 nm and the pure SDW was used for calibration and as a blank sample.

2.3 PCR amplification

2.3.1 Primers

For fragment analysis, either radioactive or fluorescent labelled primers were used, whereas unlabelled primers were applied for sequencing. The different primer-combinations and their sequences employed in analyses are described in Table 2.2.

Table 2.2: Primers/primer combination used for fragment analysis and sequencing of various genes/gene clusters.

gene/gene cluster primer-pairs used	**primer-name**	**primer-sequence (5' – 3')**
Cluster A: U31-U51	U31F	GDDATTGTCGCCCCAGKCTTAA
U31F-SR4R^{R,F}/U33R^{R} /U51R^{R,F}	U51R (F)	TCAGCCGAAAGATGGTGA
SR4F-SR33R^{R}	U33F	CATGCACTACCATCTGATCT
SR33F-U51R^{R,F,S}	U33R	AGATCAGATGGTAGTGCATG
	SR4F	TGTGACAYCCAGTCTTATCT
	SR4R	AGATAAGACTGGRTGTCACA
Cluster B: U14	U14-1F	ACATTCGCAGTDGCCGCCTA
U14-1-U14-2^{R,S}	U14-2R	TCAGACATCCAAGGAAGGA
U14-1-U14-2 variants	U14-3F (F)	TCCTTCCTTGGATGTCTGA
U14-3-U14-4^{R,F}	U14-4R	TAGGCGGCHACTGCGAATGT
	U14-2.1	TCAGACATCCAAGGAAGGAATARGC
	U14-2.1a	TCAGACATCCAAGGAAGGAATAAGC
	U14-2.1b	TCAGACATCCAAGGAAGGAATAGGC
	U14-2.2	TCAGACATCCAAGGAAGGAAAAARGC
	U14-2.2a	TCAGACATCCAAGGAAGGAAAAAAGC
	U14-2.2b	TCAGACATCCAAGGAAGGAAAAAGGC
	U14-2.3	TCAGACATCCAAGGAAGGAGCAAAAA
	U14-2.3a	TCAGACATCCAAGGAAGGAGCAAAAAC
	U14-2.3b	TCAGACATCCAAGGAAGGAGCAAAAAAC
	U14-2.3c	TCAGACATCCAAGGAAGGAGCAAAAAAAC
	U14-2.4	TCAGACATCCAAGGAAGGATARAAC
	U14-2.4a	TCAGACATCCAAGGAAGGATAAAAC
	U14-2.4b	TCAGACATCCAAGGAAGGATAGAAC
	U14-2.5	TCAGACATCCAAGGAAGGAGGAAAAAC
Cluster C: U36-U38	U36aF (F)	TGGTTGAATTTCTTRATATGAGCC
U36aF-U38R ^{R,F}	U38R	TCATGAAGCAGAACTGGC
Cluster D: U49-SR77Y	U49F (F)	GATAGGAAGTGCCGTWTGACAC
U49F-SR2dR/SR77YR^{R,F}	SR2dR	AAGATCCACAGGTTCCTATCAGTA
SR2dF-SR77YR^{R,F}	SR2dF (F)	CGTGTTTCGCTTACTGATAGGAAC
	SR77YR	TCWGACGGTAATTCCA

Cluster E: SR13-U54	SR13F (F)	GTATTTAAGTCTCTGATGAT
SR13F-U18R^{R,F}/U54R^{F}	U18R	TCAGAAACACGGACCAA
U18R-U54R^{R,F}	U18F (F)	TTGGTCCGTGTTTCTGA
	U54R	TCRGWATAGCGTATAYTGC
Cluster F: U61-SR14	U61F (F)	TACACWACCCTCTAAGAAGTTCTG
U61F-SR14R^{R,F,S}	SR14R	TCAGKGGATTGACAGAC
U61F variants-SR14R^{F,S}	U61Fc1l	ACCCTCTAAGAAGTTCTGAGCGATTACCTTTTYTTA
	U61Fc2l	ACCCTCTAAGAAGTTCTGAGCGATTACYTTTTTTTT
	U61Fc3l	ACCCTCTAAGAAGTTCTGAGCAATCATTTATTATATC
	U61Fc1s	GTTCTGAGCGATTACCTTTTYTTA
	U61Fc2s	GTTCTGAGCGATTACYTTTTTTTT
	U61Fc3s	GTTCTGAGCAATCATTTATTATATC
	U61Fc1_2	ACCCTCTAAGAAGTTCTGAGCGATTAC
Cluster G: SR29-SR30	SR29F (F)	TCAAGCTCAACAGACCGBA
SR29F-SR30R^{R,F,S}	SR30R	GGTTCGATTCTGCCAGC
SR29F-SR30R variants^{F,S}	SR30Rc1l	GGTTCGATTCTGCCAGCAAGTTGAC
	SR30Rc2l	GGTTCGATTCTGCCAGCATAGTTAC
	SR30Rc3l	GGTTCGATTCTGCCAGCAGAGTTATC
	SR30Rc3al	CTGCCAGCAGAGTTATCCTCAGAATGAAT
	SR30Rc1s	GGTTCGATTCTGCCAGCAA
	SR30Rc2s	GGTTCGATTCTGCCAGCAT
	SR30Rc3s	GGTTCGATTCTGCCAGCAG
Cluster H: U80	U80F (F)	GCATAGTTCADATG
U80F-U80R^{R,F}	U80R	TCAGATAGGAGCGAAAGAC
Cluster I: U15-SR7	U15F (F)	CGAGGCATTTGTCTGGAG
U15F-SR7R ^{R,F}	SR7R	TGAGWATGAGTAGGAGG
Cluster J: SR37-SRR80	SR37F (F)	TGGACTAGAGTTTCTGATCTGGG
SR37F-SR22R^{R,F}/SR23R^{F} /SR80 SR22F-SR23R^{R,F}	SR22R	CTCACCGAAATCCGCTAAGAT
	SR22F (F)	ATCTTAGCGGATTTCGGTGAG
	SR23R	CTCAGTGGAARGAGAAGTCGCT
	SR80R	GCATTTCCAGGATCAAW
Cluster M: SR66-Ath119	SR66F (F)	GATGGCATGWWATCTTTGAGACCTGA
SR66F-Ath119R1^{F}/R2^{F,S}	Ath119R1	CCCAGTGCAWACTTCATCATCT
Ath119bF-Ath119R2	Ath119R2	CTTTCTAGGCTGCAWTATGCATC
	Ath119bF	AGATGATGAAGTWTGCACTGGG
Cluster N: SR114-SR85	SR114F	TTGTCCGTACCATCTGA
SR114F-SR115R/SR85R^{F}	SR115R	ASCTCTCAAAGTTTGATGGTA
SR115F-SR85R^{F}	SR115F	TACCATCAAACTTTGAGAGST
	SR85R (F)	ATGTAAGGGCTTTTGA

^{R} = radioactive labeled genotyping; ^{F} = fluorescence labeled genotyping; ^{S} = sequencing; without superscript letters = tested for PCR amplification only; (F) = fluorescence labeled primer; F = forward primer; R = reverse primer. Note that only the forward primers were used for radioactive labeling.

2.3.2 PCR conditions:

Reactions were carried out in a GeneAmp®-PCR-System 2700 (Applied Biosystems) using the amplification programmes summarized in the Table 2.3Table 2.4 and Table 2.5.

Table 2.3: Standard PCR amplification profile.

cycles	temperature (° C)	time
1	95	5 min
35	95 55 72	30 sec 30 sec 45 sec
1	4	store

Table 2.4: Stringent PCR amplification profile used for samples with putative artefacts.

cycles	temperature (° C)	time
1	95	5 min
7	95 65-60 touchdown 72	30 sec 30 sec 45 sec
30	95 60 72	30 sec 30 sec 45 sec
1	4	store

Table 2.5: PCR amplification profile used for sequencing.

cycles	temperature (° C)	time
40	96 50 60	10 sec 5 sec 4 min
1	4	store

2.4 Radioactive labelled fragment analysis

2.4.1 Primer γ^{33}phosphate-end-labelling

Primers were labelled with ^{33}phosphate using a T4 polynucleotide kinase (PNK) which catalyzes the transfer of the γ^{33}phosphate of ATP to the 5' hydroxyl terminus of the primer sequence.

For 30 samples, forward primers were end-labelled in a volume of 15 µl, containing 3 µl 100 µM primer, 1.5 µl T4 PNK (Promega), 3 µl γ^{33}phosphate containing ATP (Amersham Pharmacia Biotech UK Ltd), 1.5 µl 10x PNK-buffer (20 mM Tris-HCl pH 7.5; 25mM KCl, 2mM DTT, 0.1mM EDTA, 0.1µM ATP and 50% (v/v) glycerol) and 6 µl water by incubation of 1 hour at 37 ° C and 15 minutes at 75 °C.

2.4.2 PCR amplification procedure

Polymerase chain reactions (PCRs) were performed in a volume of 10 µl, containing 1 µl dNTP mix (2 µM of each dNTP) (Roche), 1 µl 10x Taq-polymerase buffer, 0.5 µl γ^{33}phosphate end-labelled primer, 0.1 µl 100 µM reverse primer, 0.1 µl Taq-polymerase (5U/µl) (Roche), 6.3 µl water and 1 µl of template (approximately 20 ng/µl). Reactions were carried out using the standard PCR amplification profile (Table 2.3).

2.4.3 Polyacrylamide gel electrophoresis

2.4.3.1 Preparing the gel

A pair of glass plates, one with and one without a notch (upper and lower plate, respectively), was cleaned with distilled water and 70 % ethanol and the upper one was coated with Repel silane (0.2% v/v dimethyldichlorosilane in 1,1,1-trichloroethane) (BDH) on one side to prevent the gel from sticking. The clearance is needed for putting the comb into the gel. The washed spacers, thin plastic stripes, were placed on both long sides of the lower plate. After the upper plate was put onto the lower one with its coated side facing lower plate, they were clamped into a gel cassette. The gap between the two plates opposite the clearance was sealed by the rubber band of the cassette's bottom

compartment (pouring equipment). The rubber band contains an adaptor connecting the space between the two plates with the syringe necessary for pouring.

75ml of the prepared 6 % gel mixture (acrylamide - bisacrylamide (19:1), 7 M urea, 1x TBE) (Severn Biotech Ltd.) were transferred to a beaker, 50µl Temed (N, N, N', N'-tetramethylethylenediamine) (Sigma) and 550µl of 10% APS (ammoniumpersulfate) (Sigma) were added and briefly stirred. The mixture was taken up with a 25 ml pouring syringe which was then plugged to the adaptor of the pouring equipment. The gel was poured by pressing hard to avoid air bubbles. The flat side of the comb was inserted 1 to 2 cm into the plate's interspace of the notch and the gel was left to polymerize for about 2 hours. The pouring equipment was removed and the gel was put into the gel-chamber. After the gel chambers were filled up to the mark with 1x TBE buffer, the comb was removed and the gel was preheated to 40° C to 50° C by applying 90 W for about 1 hour. The comb was reinserted; just deep enough for the teeth to touch the gel and the samples were loaded.

1 µl 10x loading buffer (95% formamide, 0.05% bromophenol blue, 0.05% xylene cyanol, 20 mM EDTA) were added to each amplification product and then denatured for 10 minutes at 94° C. Six µl of each sample were loaded and after gel-electrophoresis for about 3 to 4 hours at 90 W the gel was dried for 2 hours at 80° C using a vacuum dryer (Biorad). In a dark room, the dried gel was transferred to a film (X-OMATS 100, Kodak) which was developed after 2 to 5 days incubation. The film was surveyed for (i) amplification success and (ii) fragment-length variation between the samples/species.

2.5 Fluorescence labelled fragment analysis

2.5.1 PCR amplification procedure

Standard and stringent PCR amplification (Table 2.3 and Table 2.4) profiles were used and reactions were carried out as described in section 2.4.2, with the following exceptions: instead of the radioactive labelled primer, 0.1µl 6-FAM fluorescent labelled primer (100 µM) (Operon), usually the forward one (Table 2.2), was used and 0.1 µl BSA (Bovine serum albumin; 20 mg/ml) (Fermentas) and 0.2 - 0.3 µl $MgCl_2$ (50 mM) (Bioline) were each added to the reaction mix.

2.5.2 Preparation of PCR products for ABI 3730 analysis

1 ul of a 1:5 diluted PCR product was combined with 9 µl mixture composed of 8.92 µl Hi-Di^{TM} formamide (Applied Biosystems) and 0.08 µl size standards (GenScan ROX 500 and ROX 1000, respectively) (Applied Biosystems). This mixture was then denatured at 95 °C for 3 minutes prior to loading on an automatic sequencer ABI 3730. Raw data were collected, aligned with the internal size standard and scored using Genemapper 4.0 analysis software (alternatively peakscan, both Applied Biosystems).

2.6 Sequencing

2.6.1 PCR amplification procedure

Samples were amplified as described in section 2.5.1 except the reactions were performed in a volume of 20µl (double volume of each solution) and unlabelled primer were used.

2.6.2 Cloning

2.6.2.1 Purification

Following PCR, for some primer combinations, 7 to 10 samples of the same species were pooled, while for others only PCR amplications of the same individual were combined. After pooling, PCR products were purified using a Wizard® SV Gel and PCR Clean-Up System (Promega; for procedure see DNA purification manual 9-34 (http://www.promega.com/paguide/chap9.htm#title10). 5 µl of the purified PCR products were checked in a 1.5% agarose gel.

2.6.2.2 Ligation:

The purified PCR-fragments were ligated into a pGEM®-T vector using the pGEM®-T Easy Vector System (Promega). The molar ratio between vector and insert should be between 1:3 and 3:1. The appropriate amount of insert was calculated using the equation from the manual (pGEM®-T and pGEM®-T Easy Vector Systems manual p. 13).

10 µl ligation reactions containing, 1 to 3.5 µl purified PCR product, 5 µl 2x rapid ligation buffer (Promega), 0.5 µl pGMT®-T vector (50 ng/µl) and 1 µl T4 ligase (3 U/µl) and SDW were mixed together in a 0.5 ml eppendorf tube and incubated either at room temperature for 1 hour or at 4 C o/n. In addition to the samples, control reactions with 2 µl Control Insert DNA (positive control) and without an insert DNA (negative control) were performed. Ligations which resulted in none or only few colonies were repeated with different vector:insert ratios.

2.6.2.3 Electrotransformation

The pGEM®-T vector containing the insert was transferred into electro-competent *Escherichia coli* cells via electroporation.

ElectroMAX DH10B^{TM} cells (invitrogen), normally stored at -80° C, were defrosted for several minutes on ice. The ligation mix (section…) was diluted 1:5 and 1 µl was mixed with 20µl of DH10B cells and carefully transferred into an ice-cold electroporation cuevette (Fisher Scientific). The cuevette was dried using a paper towel before it was put into the holder of the electro impulse apparatus (Biorad). After a 1.6 kV impulse was applied, the cells were resuspended in 1 ml of SOC (Super optimal broth with catabolite repression) medium (2% bactotryptone, 0.5% bacto yeast extract, 8.56 mM NaCl, 2.5 mM KCL, 10 mM $MgCl_2$, 20 mM glucose), transferred into a 15 ml plastic tube and shaken at about 200 rpm at 37° C for 1 hour. 20 µl and 100 µl respectively, were transferred on X-LBA-plates (Lauria Bertani (LB) plates (1% bactotryptone, 0.5% bacto yeast extract, 1% NaCl, 1.5% agar), 100µl 2% X-gal (in dimethylformamide), 10µl 100 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside), 10µl ampiclilin (100mg/ml)) and evenly distributed using a glass triangle rod, sterilised by flaming. This step was performed on a flow bench to avoid contamination. The plates were incubated at 37° C o/n (16 to 24 hours).

2.6.2.4 Multiscreen Plasmid Minipreparation (Millipore)

Following blue/white selection, white colonies were picked and inoculated into 1 ml of 2x LBA broth (2% bactotryptone, 1% bacto yeast extract, 2% NaCl, 100µl/l ampiclilin (100mg/ml)) in 96-well deep blocks (plates) which were then covered with a gas permeable sheet (AB Gene) and shaken at 300 rpm at 37° C for 24 hours. Plates were spun down at 3000 for 5 minutes using a centrifuge 5810R (Eppendorf), the supernatant decanted and any residual media removed by inverting the plate on a paper towel. The pellets were fully resuspended in 80 µl of Solution I (30 mM glucose, 15 mM Tris-HCL (pH 8.0), 60 µg/ml RNase A; stored 4° C). 80 µl of Solution II (0.2 M NaOH, 1 % SDS (sodium dodecyl sulphate; fresh made) were added, vortexed for 1 minute and left for 2 minutes at room temperature (RT). After 80 µl of Solution III (3.6 M potassium, 6 M

acetate) were added and vortexed for 1 minute, 130 µl lysate were transferred to each well of the Multiscreen MANANLY clearing plate (Millipore). 160 µl binding solution were added to each well of the Multiscreen MAFBNOB binding plate. Whereas the binding plate was placed in the base of the manifold, the clearing plate was placed on the top. 10'' Hg vacuum were applied to manifold for 3 minutes drawing the lysate through the binding plate into the wells. The binding plate was placed on top of the manifold and the lysate was mixed with the binding buffer by pipetting 3 times. To collect the waste, an inverted lid was placed in the base of the manifold and full vacuum was applied for 1 minute. After the waste was disposed, 200 µl of 70 % ethanol were added to each well and full vacuum was applied for 1 minute. Another 200 µl of 70% ethanol were added, full vacuum was applied for 3 minutes and the plate was plotted on a paper towel. The binding plate was placed on a microtitre plate using Millipore alignment frames (MACF09604) and spun at 3000 rpm for 10 minutes. After the membranes were air dried for 10 minutes, 75 µl SDW were added to each well, the binding plate was placed on a new microtitre plate and the plasmid was eluted by spinning at 3000 rpm for 5 minutes.

2.6.2.5 Insert size determination

Insert sizes were estimated by digestion of the plasmid with *Eco*RI restriction enzyme. 10 µl plasmid were digested by adding 1.5 µl 10x restriction buffer, 0.5 µl EcoRI (12U/µl) (Promega) and 3.5 µl SDW and incubation at 37° C for 1 hour. Samples were run on 1.5% agarose gels and plasmids containing an insert were taken for sequencing.

2.6.3 PCR-sequencing

Plasmid DNA was sequenced using the automated fluorescent sequencer ABI 3730 (PE, Applied Biosystems), Sequencing reactions were carried out using 3 µl of plasmid DNA, 3.2 pmoles of primer (M13 reverse or M13 forward) and 0.5 µl Big Dye reaction mix (Applied Biosystems) in a final volume of 10 µl. The reaction was carried out in a GeneAmp®-PCR-System 2700 (Applied Biosystems) using the PCR amplification for sequencing (Table 2.5).

2.6.4 Direct colony PCR sequencing

Sequencing of cloned products was also carried out by direct colony sequencing. White colonies were picked, transferred into 10 µl SDW, denatured at 96° C for 5 minutes and spun down. 6 µl were used as template DNA and sequenced as described in section 2.6.3.

2.6.5 Direct sequencing from PCR products

Following PCR, 1 µl of ExoSAP-IT® exonuclease (USB) was added to 5 µl of PCR and incubated at 37° C for 15 minutes followed by 80° C for 15 minutes. These were then subjected to sequencing as described in section 2.6.3 with the exception that one of the primers used for PCR amplification was applied.

2.6.6 Precipitation of sequence reactions

After performing PCR, the unincorporated dye terminators were removed. The reaction plate was briefly spun down and 2.5 µl of 125 mM EDTA followed by 30 µl of 100 % ethanol were added to each well. After sealing the plate with an aluminium foil, the plate was inverted 4 times, incubated for 15 minutes at RT and centrifuged at 2000 to 3000 x g at 4° C for 30 minutes. The supernatant was removed by inverting the plates on a paper towel and centrifuged at 185 x g. 30 µl of 70% ethanol were added to each well and the plate was spun at 1650 x g at 4° C for 15 minutes. The supernatant was removed by repeating the centrifugation step at 185 x g for 1 minute. The samples were covered with aluminium foil and stored at 4° C.

2.6.7 Preparing reactions for ABI 3730 sequencing

Before the samples could be loaded on an automatic sequencer ABI 3730 they were resuspended in injection buffer. Raw sequence data were collected examined using Sequencher 4.7 (Gene Codes Corporation).

The analyses of the molecular data will be described in their respective chapters.

Chapter 3: Developing molecular marker systems based on snoRNA genes

3.1 Introduction

snoRNAs do not code for proteins but contain conserved regions which allow the development of PCR-based markers (Brown *et al.*, 2003a). In addition, numerous snoRNA gene sequences are available from different taxa (e.g. Plant snoRNA database, TAIR and Genebank) providing a breadth of sequence information on which to base potential markers for DNA barcoding and phylogenetic studies. Phylogenetic studies are usually restricted to certain species groups (e.g. species complexes, genera and families) and, therefore, it is possible to choose the most suitable marker for the group under investigation. A marker for DNA barcoding should be able to identify species within much larger groups like land plants or animals (Chase *et al.*, 2005; Hebert & Gregory, 2005). While the mitochondrial gene CO1 is an effective DNA barcode region in animals (Hebert *et al.*, 2003), in land plants, although several regions have been proposed (e.g. Kress *et al.*, 2005; Chase *et al.*, 2007), there has been no agreement on which region(s) should be used (Pennisi, 2007; Kane & Cronk, 2008). A key step in the development of new markers is primer design. Of particular interest for putative primer sites are the antisense elements of box C/D snoRNA genes (10 to 21 nucleotides long) which base-pair with specific ribosomal RNA target regions (Brown *et al.*, 2003a). Most of these elements do not only have the desired primer length, at least together with their adjacent box D or D' sequence, but their sequences differ between different snoRNA genes. Some genes contain two antisense elements and primers can be designed to amplify only the intragenic sequence (Figure 3.1B). Most of the snoRNA genes, however, contain only one antisense element. Fortunately, in plants the majority of the snoRNA genes are found in gene clusters, some with conserved gene order (Brown *et al.*, 2001). Thus, the antisense elements of different genes of the same cluster can provide the two primer sites required. Therefore, it is possible to design primers for the amplification of both intra- and intergenic sequences (Figure 3.1A and B). Due to the short length of snoRNA genes, however, it is desirable to design primers amplifying two or more genes.

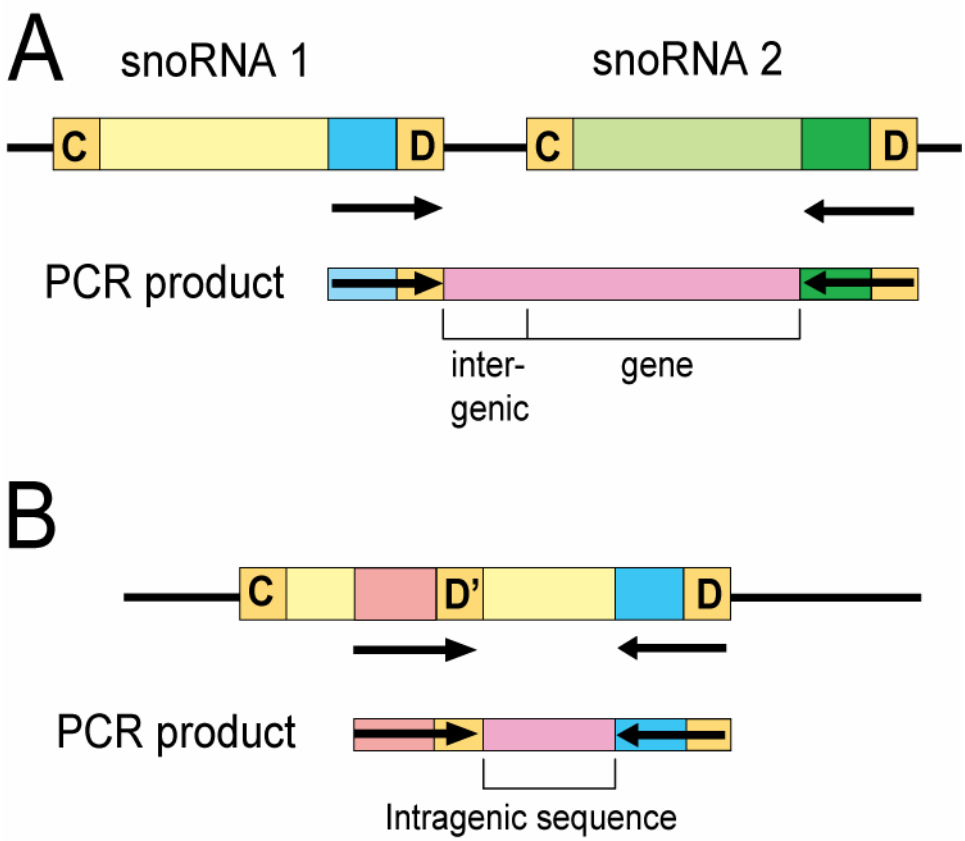


Figure 3.1: Structure of box C/D snoRNA genes and their conserved regions used for primer design. A: snoRNA genes containing one antisense element. B: snoRNA gene with two putative primer sites. Black line = intergenic region; C, D and D' = conserved boxes; coloured boxes upstream of the boxes = antisense elements; black arrows = primer sites.

Although box C/D snoRNA genes are the first choice for the design of primers, box H/ACA snoRNA genes might also contain conserved sequences useful as primer sites. Ideally, universal primers are usually 15 to 30 nucleotides long and amplify orthologous DNA sequences in a wide range of related species. The many snoRNA sequences available in data bases were used for generating sequence alignments containing homologues of several species. Usually, I refer to homologues rather than to orthologues and paralogues because in most cases it cannot be determined if a homologous sequence represents an orthologous or paralogous gene. Alignments were examined for conserved regions suitable for primer sites and primers were designed, characterized and tested initially using a computational or virtual PCR method – electronic PCR (see below).

3.1.1 BLAST searches and sequence libraries

BLAST compares sequences to sequence databases and directly delivers alignments which show the optimal local similarity measured by the maximal segment pair (MSP) score (Altschul *et al.*, 1990). In other words, the programme calculates the statistical significance of the matches and was, thus, used for finding *Arabidopsis thaliana* homologous snoRNA genes and gene clusters in other species.

As whole genome sequencing is very expensive and currently impracticable, especially for organisms with large genome sizes, most of the sequences available are obtained from expressed sequence tags (EST) libraries. ESTs are generated from messenger RNA (mRNA) representing the actively expressed genes of a cell/organism. The mRNAs are transcribed into double stranded cDNA which can be cloned and sequenced. The clones are sequenced randomly to generate the EST sequences. While genes with high expression rates are sequenced multiple times, rare transcripts are under-represented resulting in sampling biased EST libraries containing not more than 60 % of the genes (Bonaldo *et al.*, 1996). Furthermore, EST libraries are usually redundant because the cloned sequences can be of partial (e.g. due to mRNA processing) or full length. Additionally, the ESTs within a library are error prone because they are usually sequenced only once (Nagaraj *et al.*, 2007). Despite these drawbacks, EST libraries are relatively cheap to generate, are available for myriad organisms and are highly useful for comparative genomics, especially for gene discovery and characterisation.

3.1.2 Electronic PCR (ePCR)

Whether or not a particular primer pair will amplify the desired region in experimental PCR depends on various poorly understood factors (e.g. interactions between the compounds of the PCR reaction mix) and, therefore, it is not possible to model the PCR process in detail (Bangham, 1991). However, a very useful application for testing the designed primers, which has been successfully used for the determination of the genomic location of markers, the examination of the uniqueness of primers and the prediction of possible fragment sizes (Thongjuea *et al.*, 2009; Yonemaru *et al.*, 2009; Hyten *et al.*, 2010; You *et al.*, 2010), is e-PCR. Similar to BLAST searches but without possible false

positive matches due to sequence similarities to pseudogenes and related gene family members, this programme searches sequence databases for sequences matching the primer pair (Schuler, 1997). By comparing the newly designed primers to genomic sequences (e.g. in the *Arabidopsis thaliana* genome), the programme does not only check the correct order of a primer pair and predict the length of the putative PCR fragments, but might also discover multiple amplification sites within the genome. Thus, primers with unintentional multiple matches can be discarded before using them in experiments (Rotmistrovsky *et al.*, 2004). Furthermore, the location and the gene reference, if existent, of a putative product is given as well. Additionally, any errors in designing the primers (e.g. missing a base) are discovered easily without amplification. Thus, e-PCR is a computational method used to identify sequence tagged sites (STS), unique sequences within the genome defined by a primer pair and the expected product size, within a DNA sequence (Olson *et al.*, 1989; Schuler, 1997). Short sequences (words) from the 3' end of each primer are stored in a sorted hash table. Longer sequence lengths used for hashing accelerate the search by reducing the number of matches which has to be investigated. To increase the sensitivity of the search overlapping discontinuous words allowing mismatches and gaps in the alignment between primer and sequence are introduced. A match is reported if both primers have their right orientation, the number of allowed mismatches and gaps is not exceeded and the size of the STS is within the expected range. While forward e-PCR searches STS databases with sequences, STS are used to search sequence databases in reverse e-PCR searches (Rotmistrovsky *et al.*, 2004).

3.2 Material and methods

Arabidopsis thaliana snoRNA gene/gene cluster sequences, identified by Marker et al. (2002) and Brown et al. (unpublished data), were provided in FASTA format.

3.2.1 BLAST searches

In an initial screen, sequence alignments that were available (e.g. Plant snoRNA database - http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home) were examined and snoRNAs

that did not have conserved sequences of >18 nt in length were discarded. It should be noted, that all alignments shown in this thesis were updated and that some primers were designed using a subset of the sequences available at the time of designing. Thus, some primers might not appear to fit the alignments. Furthermore, BLAST searches conducted earlier were also performed against the *Brassica* database (no longer available) on the TAIR website.

The remaining snoRNA gene/gene cluster sequence (FASTA format) was entered in the interface and web based nBlast searches were performed against the nucleotide collection (nr/nt) and ESTs (NCBI server: http://www.ncbi.nlm.nih.gov/) and Green plant GB genomic (DNA) and experimental cDNA/EST (DNA) (TAIR server: http://www.arabidopsis.org/). Default settings were used except on the NCBI website where the 'Somewhat similar sequences (BLASTN)' option was chosen (default: Highly similar sequences (megablast)). Adjacent genes in a gene cluster were examined using either the sequences obtained from BLAST searches of single genes and neighbouring genes or by BLAST search of the entire cluster sequence.

Putative homologous gene sequences from different species were examined before copying them (in FASTA format) to a single file. Gene cluster conservation was investigated by looking for and examining neighbouring snoRNA genes in the EST and genomic sequences and comparing the gene order to other species. Homologous gene sequences were assembled using the clustalW multiple alignment (Thompson *et al.*, 1994) as incorporated in BioEdit version 7.0.9.0 (Hall, 1999). By aligning each pair of sequences separately, Clustal W calculates a distance matrix from which a guide tree is generated. According to the branching order of the tree, sequences are progressively aligned using different weight matrices to optimize gap penalties (Thompson *et al.*, 1994). Conserved regions, usually the antisense elements and box D or D', were identified and used to design primers. Because of the limited possibilities of putative primer sites, primers have to be designed by hand rather than by programmes like primer3 (http://frodo.wi.mit.edu/primer3/input.htm).

3.2.2 Primer characterization and reverse ePCR

Primers were characterised (e.g. by length, basic melting temperature (TM), GC content; http://insilico.ehu.es/tm.php) and suitable primer combinations were tested virtually using ePCR (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/). Melting temperature (TM) was calculated using the thermodynamic nearest-neighbour model (SantaLucia, 1998), which takes not only the melting temperature of the single bases but also their sequence (neighbouring bases are influenced by each other) into account. Estimations were executed using the default settings (c (primer) = 200 nM; c (salt) = 50 mM; c (Mg^{2+}) = 0 mM) and as wobbles cannot be used for calculations they were substituted for their respective bases and the range of TM for each primer was determined.

For testing the designed primer combination, the reverse e-PCR procedure on the e-PCR Web Server (http://www.ncbi.nlm.nih.gov/projects/e-pcr/reverse.cgi) was used. To run a reverse e-PCR, a sequence database was selected (e.g. *Arabidopsis thaliana* genome ref_assembly 8.1 database) and the names and sequences of up to five chosen primer pairs were entered in the table. While the length of the words (W) and the number of the discontinued words (F) used for hashing is fixed, the number of gaps (G, 0-2), mismatches (N, 0-2), the expected length of the e-PCR product (0-350 default) and size deviation of the expected sequence length (M) can be chosen. Searches were conducted for all primer combinations with G=2, N=2, M=1000 and expected product length as default against the *Arabidopsis thaliana* genome ref_assembly 8.1 and transcriptome snapshot 2009/01/06 as well as the *Oryza sativa* genome ref_assembly 4.1 and transcriptome snapshot 2009/01/06 databases. However, primers might contain wobble bases (e.g. K equates A or T) which are causing mismatches by default (personal observation) and might result in under-representation of matches found by each primer combination. Therefore, the wobble bases were exchanged with the corresponding bases found in *Arabidopsis thaliana* ("diswobbled" *A. thaliana* primers) because testing every possible sequence variation, caused by these wobbles, would be too time consuming. Additionally, for the primer combinations designed for cluster A, every possible primer sequence (completely refined sequences) pair was tested.

3.3 Results

Nucleotide BLAST searches using whole *Arabidopsis thaliana* snoRNA gene sequences identified different numbers of homologous sequences in various plant species for all genes (Table 3.1). While genes of certain gene clusters are present in many species, other gene clusters harbour genes with homologous sequences found in relatively few species. For instance, snoR37 and snoR80 (cluster D, Table 3.1) are found in twenty-one and twenty species, respectively, while genes 424, 502 and snoR95 (cluster A, Table 3.1) could be identified in only six, five and three species, respectively. Furthermore, the number of homologous sequences found varies greatly between different genes within some clusters. For example, within cluster C (Table 3.1) snoR66 and 119b are present in 19 and 18 species, respectively, whereas gene 382 could only be found in 4 species. Another example is cluster B (Table 3.1), in which snoACA-1 could only be detected in two species (including *A. thaliana*), but snoR68, 319, 122, 118a/b were found in 10, 7, 8 and 9 species, respectively. Although gene 382 and snoACA-1 are both H/ACA box snoRNA genes, which are more difficult to identify than box C/D snoRNA genes, it is very unlikely that the lack of homologues in other species is due to identification errors. Other box H/ACA genes could be identified in approximately the same number of species as box C/D genes. For instance, in cluster E (Table 3.1) box C/D snoRNA genes snoR114 and snoR115 were found in 18 and 10 species, respectively, and snoR85a/b was present in 19 species. From the conserved sequences within each gene, putative primer sites were designed for one to three regions. Primers were designed for all genes with the exception of snoACA-1 where only two gene sequences were available. For most genes only one primer per conserved region, consisting of the complete conserved sequence, was designed. For two genes (snoR66 and 119b) alternative primers (sequences in italics, Table 3.1) of different length and characteristics were designed using either a part of the conserved region (snoR66) or 5’ end extension of the first 8 bases of the conserved region. The length of the putative primers ranges from 15 bp (119b primer 2) to 26 bp (snoR66 primer 1), the GC content from 33.3 % (122 primer 2, 118 primer 3, snoR80 prime 1 and snoR115) to 66.7 % (122 primer 1) and the basic melting temperature (TM), depending on length and GC content, from 41.7 °C to 63.8 °C (Table 3.1).

Table 3.1: Primers designed for various snoRNA genes and gene clusters. Homologous sequences of each gene were aligned to identify conserved regions which were used to design primers. In two cases alternative (more conserved part of the putative primer site) sequences (in italics) were chosen as well. Please note that the primer sequences are all written in their forward 5' 3' direction. To obtain a suitable primer pair the downstream primer has to be translated to its complementary sequence.

snoRNA gene	Type	No of homologues	No of species	No of possible primer sites	Putative primer sequences (forward 5'-3')	length (bp)	TM (C)	GC (%)
424-502-snoR95 cluster (cluster A)								
424	H/ACA	6	6	1	ATAGCCCCTTGCWWCTT	17	54.2-55	47.1
502	C/D	5	4	1	CTTCAAAGTTCTCTGA	16	41.7	37.5
snoR95	H/ACA	3	3	1	CTAYACGAGCATGGTGC	17	49.2-52.5	52.9
snoACA1-snoR68-319-122-118a-118b cluster (cluster B)								
snoACA-1	H/ACA	2	2	-	-	-	-	-
snoR68	C/D	10	10	1	TGGTTCGTATTCVCTGAGCA	20	53.7-56.3	45
319	H/ACA	7	7	1	CCAAGTTTRCCTTCGDAWAT	20	50.1-55	35
122	H/ACA	8	8	2	GCGAAGGDCCCAGCAGRG	18	57-62.3	66.7
					TGAGDCYTCTCTAACAAT	18	44.1-49	33.3
118a/b	H/ACA	11	9	3	GTGTGTATCGGCKTWGTGC	19	56-58.1	52.6
					AGRTGGGCAGTTGTGHTTCA	20	53.9-58.7	45
					TCAACAATCATYTTCCCYACA	21	48.1-52.4	33.3
382-snoR66-119b cluster (cluster C)								
382	H/ACA	4	4	1	GCARGGGCGYTGAGTCGCTT	20	60.2-63.8	60
snoR66	C/D	24	19	1	GATGGCATGWWATCTTTGAGACCTGA	26	60.5-61	42.3
					TGATGGCATGAAATCTTTG	19	48.5	36.8
119b	C/D	22	18	2	GCACTGGGCTCTGAG	15	50.5	66.7
					AGATGATGADTDTGCACTGGG	21	52.2-56.7	45.5
					GATGCATAWTGCAGCCTAGAAAG	23	55.5	43.5
snoR37-snoR22-snoR23-snoR80 cluste (cluster D)								
snoR37	C/D	26	21	2	GTGGACTAGAGTTTCHGATC	20	49.6-52.1	45
					AACCCTTGGCTGTCTGAG	18	53.6	55.6
snoR80	H/ACA	24	20	2	TTACCAATTCTGRRGGAT	18	44.7-49.4	33.3
					TTTGATCYTGAAABGCCWC	19	50.5-55	36.8
snoR114-snoR115-snoR85a-snoR85b cluster (cluster E)								
snoR114	C/D	19	18	1	TTGTCCGTACCATCTGA	17	49.2	47.1
snoR115	C/D	10	10	1	TACCATCAAACTTTGAGAGST	21	49.6-51.5	33.3
snoR85a/b	H/ACA	33	19	1	AAGGCAAYAAATTAGAGTCTCTG	23	50.6-53.2	34.8

3.3.1 Alignments of snoRNA genes and identification of putative primer sites

The generation of alignments of snoRNA sequences allowed potential genes and gene clusters to be selected for primer design. The number of homologous sequences obtained varied for different genes from 1 to 33 (Table 3.1). Low numbers of sequences reduced confidence of identifying conserved sequences of appropriate length. Similarly, once aligned, the conserved region for some snoRNAs was too short to allow primer design. These snoRNAs were not considered further in this analysis. Of the many sequences, putative candidate genes/gene clusters, mostly single copy sequences, were investigated for possible primer sites: snoR37 and snoR80 (Figure 3.2A). Ath-424, 502 and snoR95 (Figure 3.2B), snoR68, Ath-319, Ath-122 and Ath-118a/b (Figure 3.2C), snoR66 and Ath-119b (Figure 3.2D) and snoR114, snoR115 and snoR85a/b (Figure 3.2E).

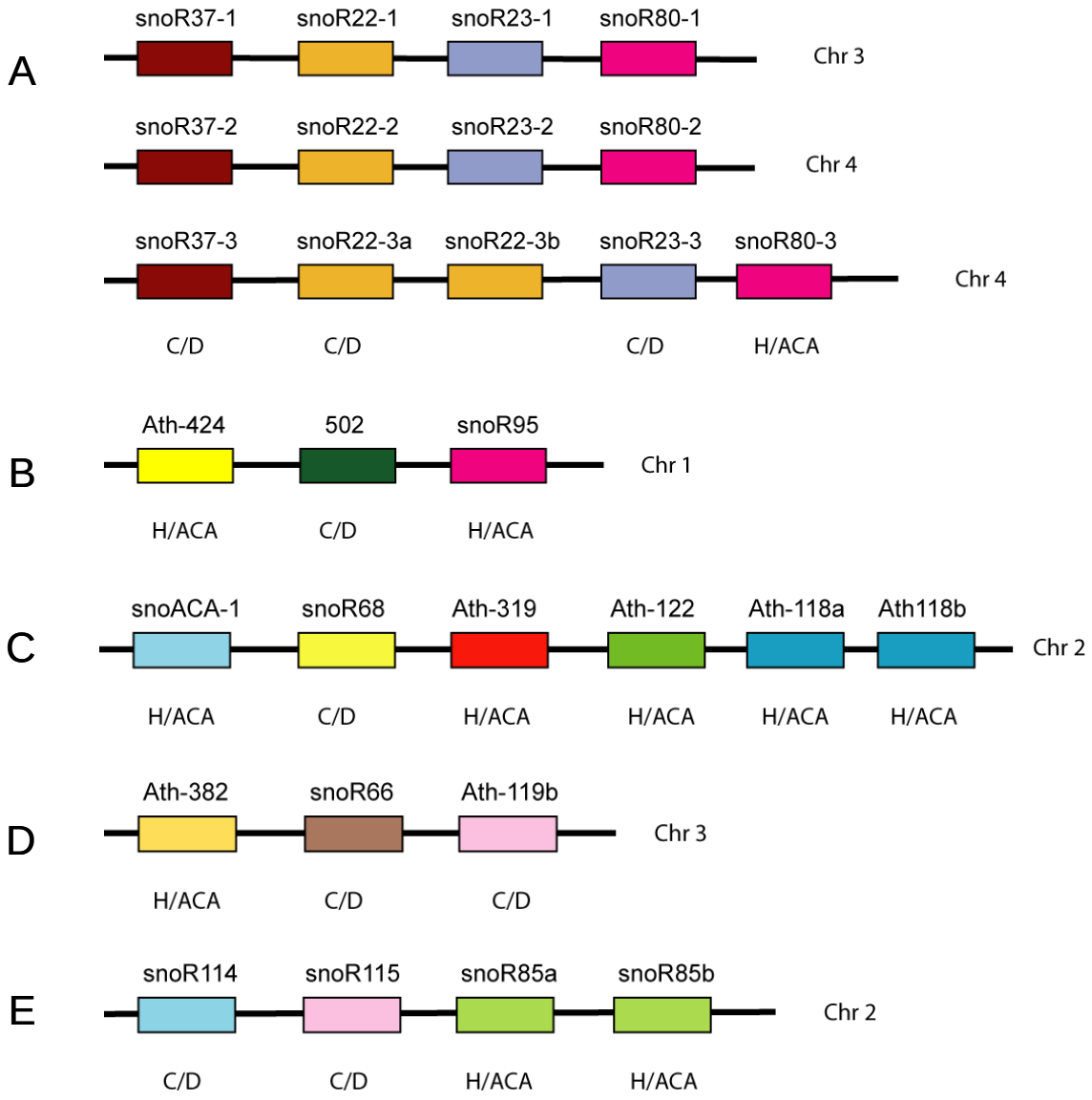


Figure 3.2: SnoRNA genes for identifying and designing primers. Eighteen snoRNA genes found in five gene clusters in *Arabidopsis thaliana* (Ath) were investigated for possible primer sites. (A) Gene cluster with three copies. (B-E) single copy gene clusters. Boxes represent gene sequences and different genes within a cluster are indicated by different colours. The names of the genes are given above the boxes, the type of the snoRNA gene (i.e. box C/D and box H/ACA genes) below the boxes and the chromosome(s) they are found on to the right of a cluster. Please note that Ath is the species abbreviation of *A. thaliana* and will be removed when discussing the genes in general.

3.3.1.1 424-502-snoR95 gene cluster (cluster A)

Homologues of the three genes, 424, 502 and snoR95 (Figure 3.2B) are aligned below and possible primer sequences are summarised in Table 1. Six H/ACA snoRNA gene 424 homologues representing six different species were identified and aligned. A putative primer site (17 bp, consensus: 5' ATAGCCCCTTGCWWCTT) at the beginning of the gene was identified. Due to the low similarity downstream of box H, only the 5' part of the gene was used in the alignment (Figure 3.3).

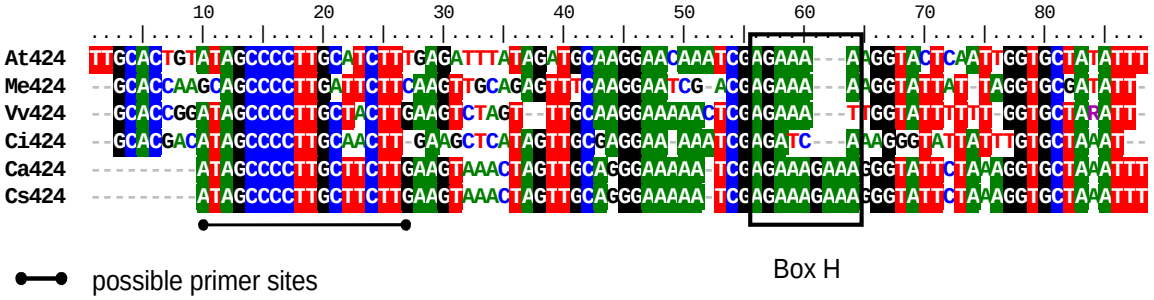


Figure 3.3: Alignment of the first 90 bp of snoRNA 424 homologous sequences from six different species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Mt - *Medicargo trunculata;* Vv – *Vitis vinifera;* Ci - *Cichorium intybus;* Ca - *Citrus aurantiifolia;* Cs - *Citrus sinensis.*

Five C/D box snoRNA gene 502 homologues (about 80 bp in length) found in four species (two homologues in *Clemone spinosa*, Clspi-502a and b) were aligned and one possible primer site (16 bp, consensus: 5' CTTCAAAGTTCTCTGA) was discovered at the end of the gene containing the antisense element and the box D (Figure 3.4).

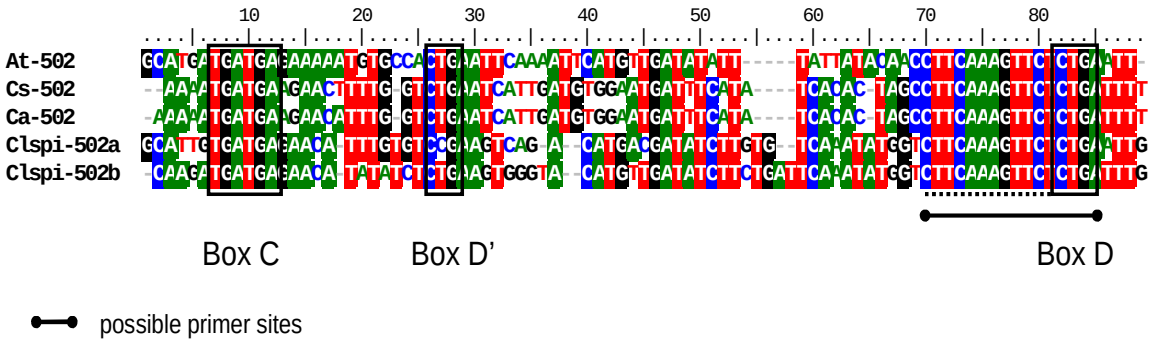


Figure 3.4: Alignment of five box C/D snoRNA gene 502 homologous sequences found in four different species. Conserved positions are shaded. Dotted line – antisense element. At – *Arabidopsis thaliana;* Cs - *Citrus sinensis;* Ca - *Citrus aurantiifolia;* Clspi – *Clemone spinosa*.

The alignment of the box H/ACA snoR95 homologues consists of only three sequences from three species. One possible primer site (17 bp, consensus: 5' CTAYACGAGCATGGTGC) was discovered which includes the box H (Figure 3.5).

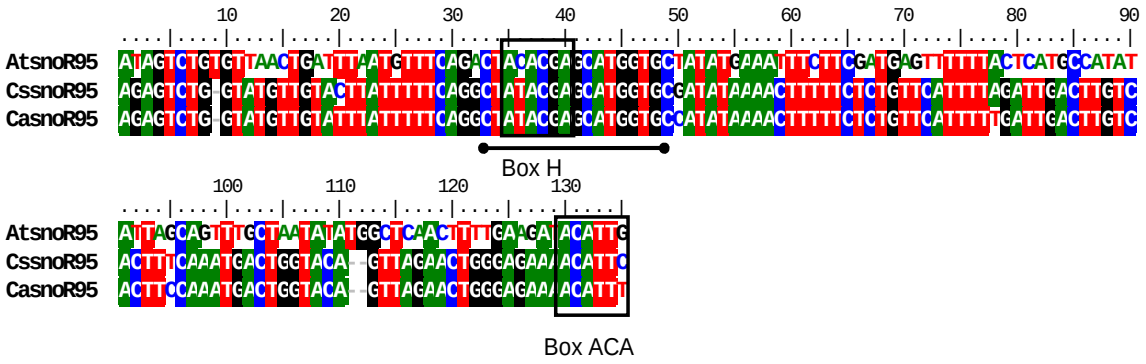


Figure 3.5: Alignment of three box H/ACA snoR95 gene homologues. Conserved positions are shaded. At – *Arabidopsis thaliana;* Cs - *Citrus sinensis;* Ca - *Citrus aurantiifolia.*

3.3.1.2 snoACA1-snoR68-319-122-118a-118b gene cluster (cluster B)

Homologues of the five snoRNA genes contained in the Arabidopsis thaliana gene cluster snoACA1-snoR68-319-122-118a-118b (Figure 3.2C) were aligned. Only one other homologue of the box H/ACA snoACA-1 gene was found in *Brassica rapa* subsp. *pekinensis* and therefore no alignment was produced for this gene. Ten box C/D snoR68 homologous sequences, each from a different species, were aligned and one possible primer site (23 bp, consensus: 5' TATTGGTTCGTATTCVCTGAGCA) was revealed in the middle of the gene, including the antisense element and the box D' (Figure 3.6).

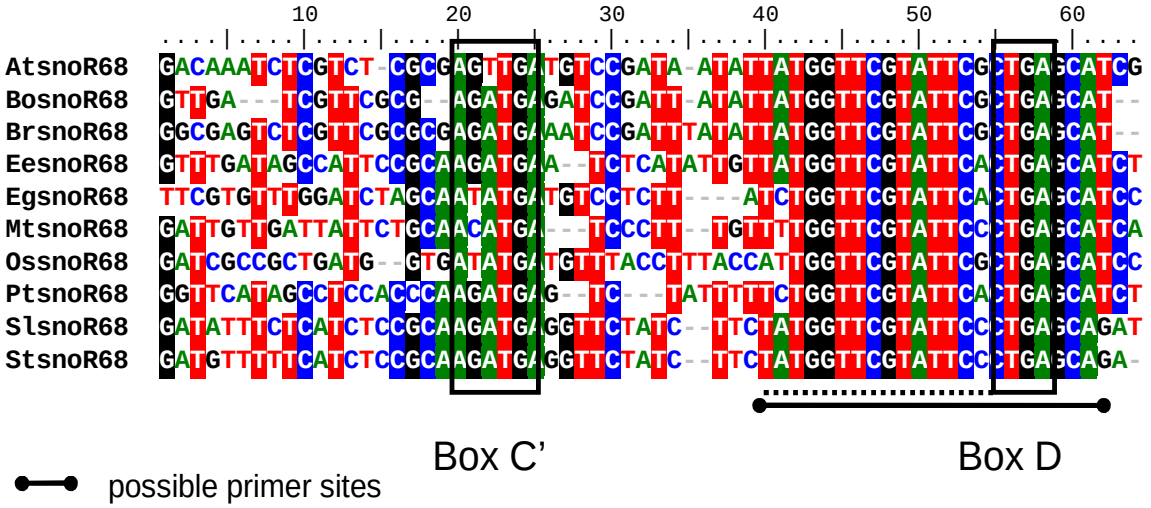


Figure 3.6: Alignment of ten box C/D snoR68 gene homologous sequences from ten different species. Conserved positions are shaded. Dotted line – antisense element. At – *Arabidopsis thaliana;* Bo - *Brassica oleraceae;* Br – *Brassica rapa;* Ee - *Euphorbia esula;* Eg - *Elaeis guineensis;* Mt - *Medicargo trunculata;* Os - *Oryza sativa;* Pt - *Populus tremula;* Sl - *Solanum lycopersicum;* St - *Solanum tuberosum.*

Seven homologous sequences of the H/ACA snoRNA gene 319, each found in a different species, were aligned. Due to low sequence similarity downstream of box H, only the first 120 bp of the *Brassica rapa* subsp. *pekinensis* and *Raphanus sativus* and 70 bp of the *Aedes aegyptii, Glycine max, Lactua sativa, Medicargo trunculata* were used in the alignment. However, one possible primer site (20 bp, consensus: 5' CCAAGTTTRCCTTCGDAWAT) was discovered upstream of box H (Figure 3.7).

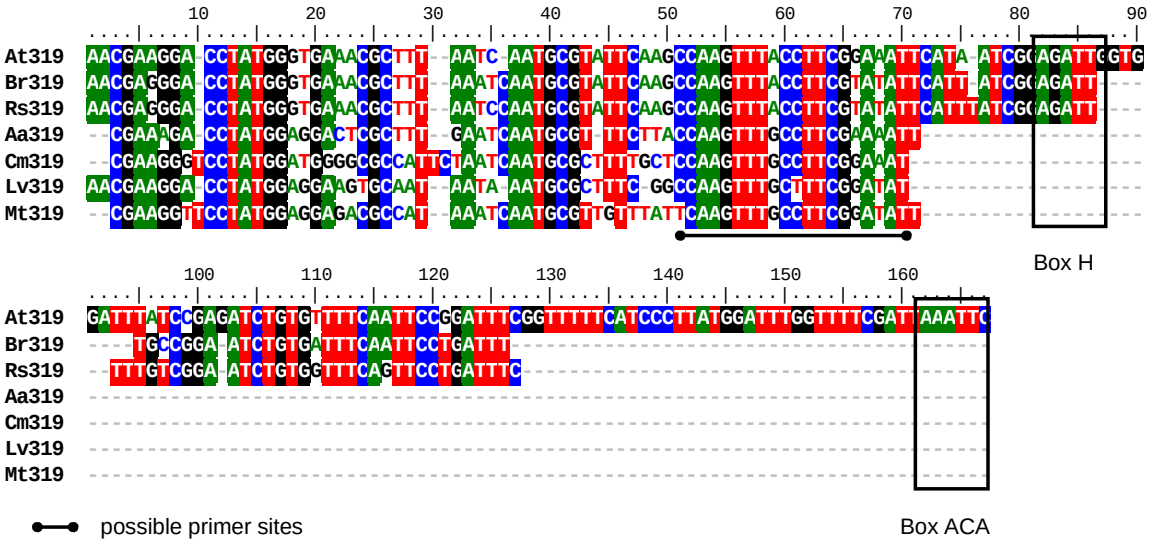


Figure 3.7: Alignment of seven box H/ACA snoRNA gene 319 homologous sequences from seven species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Rs - *Raphanus sativa;* Aa - *Acorus americanus;* Gm – *Glycine max*; Ls - *Lactua virosa;* Mt - *Medicago trunculata.*

Full-length homologues of H/ACA snoRNA gene 122 were found in *Brassica rapa* and *Raphanus sativa* and sequence homology restricted to the first part (60 to 90 bp in length) of this gene was found in five other species. These eight sequences were aligned and two putative primer sites were identified within the first 60 bp (both 18 bp, consensus 1: 5' GCGAAGGDCCCAGCAGRG, consensus 2: 5' TGAGDCYTCTCTAACAAT) (Figure 3.8).

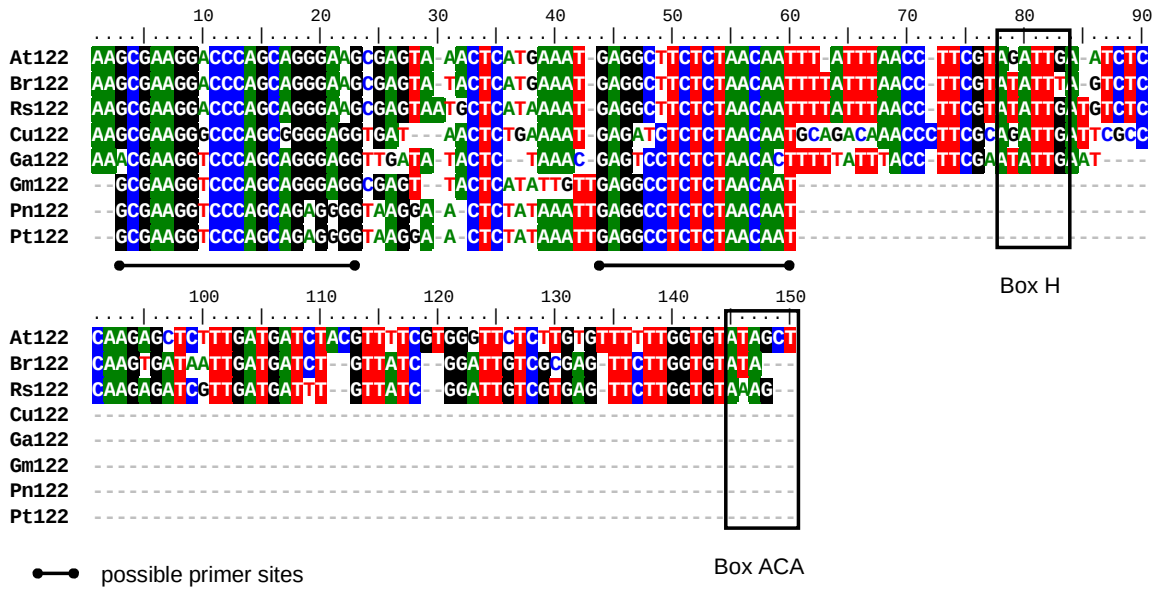


Figure 3.8: Alignment of eight box H/ACA snoRNA gene 122 homologous sequences from eight species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa* subsp. *pekinensis;* Rs - *Raphanus sativa;* Cu - *Citrus unshiu;* Ga - *Guizotia abyssinica;* Gm – *Glycine max;* Pn - *Populus nigro;* Pt - *Populus tremula.*

Three possible primer sites, one (19 bp, consensus: 5' GTGTGTATCGGCKTWGTGC) at the beginning of the gene, one (20 bp, consensus: 5' AGRTGGGCAGTTGTGHTTCA) upstream and close to box H and one (21 bp, consensus: 5' TCAACAATCATYTTCCCYACA) at the end of the gene (including box ACA), were discovered using an alignment of eleven box H/ACA snoRNA gene 118 homologous sequences (about 150 bp in length) obtained from nine species. There were two copies present in *A. thaliana* and *Raphanus sativa* (Figure 3.9).

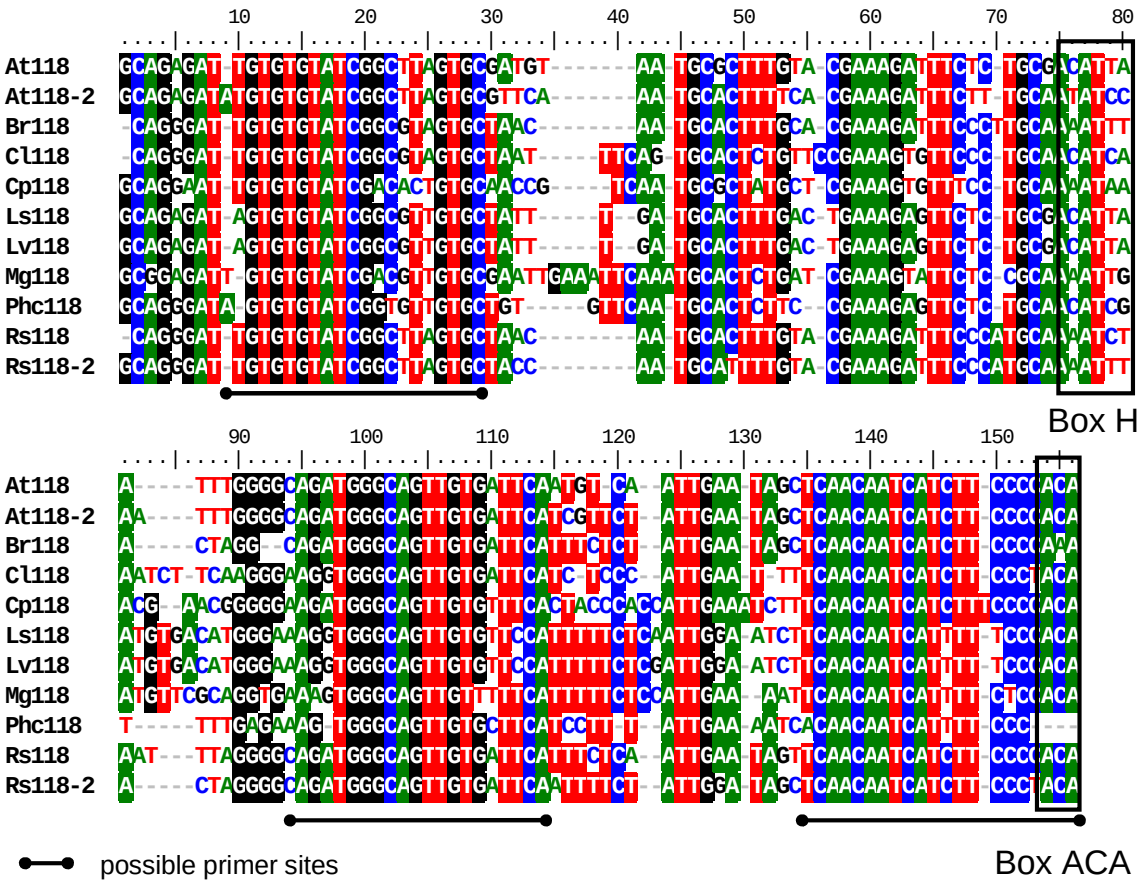


Figure 3.9: Alignment of eleven box H/ACA snoRNA gene 118 homologous sequences of nine species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa;* Cl - *Citrus lantana;* Cp - *Carica papaya;* Ls – *Lactuca saligna;* Lv – *Lactuca virosa;* Mg - *Mimulus gutatus;* Phc - *Phaseolus coccineus;* Rs - *Raphanus sativa.*

3.3.1.3 382-SnoR66-119b gene cluster (Cluster C)

Homologues of three genes in the *A. thaliana* gene cluster 382-snoR66-119b (Figure 2D) were aligned and possible primer sequences identified (Table 3.1). Homologues of the H/ACA box gene 382 (about 150 bp in length) were found in three species. Although one of the homologous sequences (Rr382) was truncated in the EST clone and missed the front half of the gene one putative primer site (20 bp, consensus: 5' GCARGGGCGYTGAGTCGCTT) it could be identified near to box H (Figure 3.10).

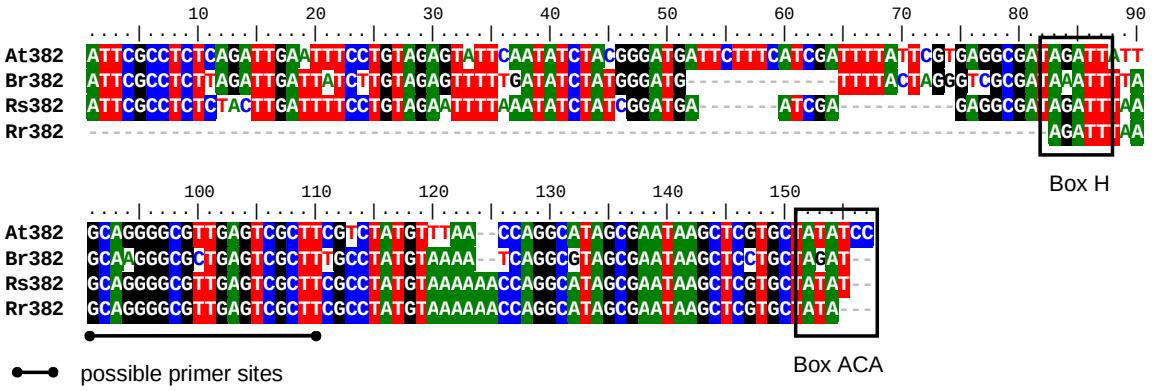


Figure 3.10: Box H/ACA snoRNA gene 382 sequence alignment of four homologous sequences found in four different species. Conserved positions are shaded. At – *Arabidopsis thaliana*; Br – *Brassica rapa* subsp. *pekinensis*; Rs – *Raphanus sativus*; Rr – *Raphanus raphanistrum* subsp. *landra*.

Twenty-four homologous C/D box SnoR66 sequences (about 90 bp in length) from eighteen different species were aligned and one putative primer site (26 bp, consensus: 5' GATGGCATGWWATCTTTGAGACCTGA) reaching from box C to box D' could be identified including the antisense element (dotted line, Figure 3.11). In most of the species examined only one snoR66 sequence was found. However, two different snoR66 sequences were observed in *Vitis vinifera* (VvsnoR66) and, surprisingly, six different sequences (OssnoR66a-d/g/h) were identified in *Oryza sativa* (Figure 3.11). As this primer site is quite long, an alternative primer sequence (19 bp; consensus: TGATGGCATGAAATCTTTG) containing only a part of the possible primer site was chosen.

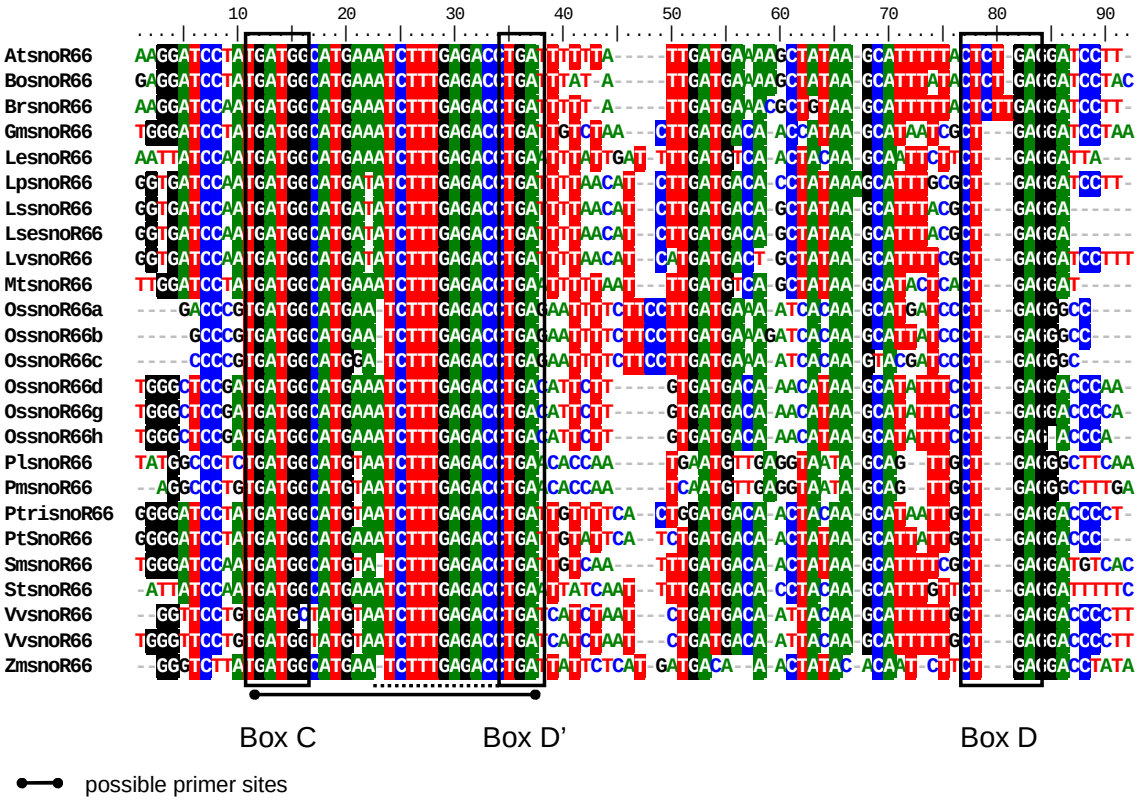


Figure 3.11: snoR66 box C/D gene sequence alignment of twenty-four genes from nineteen species. Conserved positions are shaded. Dotted line – antisense element sequence. At – *Arabidopsis thaliana*; Bo - *Brassica oleraceae;* Br – *Brassica rapa;* Gm – *Glycine max;* Le - *Lycopersicon esculentum;* Lp – *Lactua perennis;* Ls - *Lactua serriola;* Lv – *Lactua virosa;* Ht - *Helianthus tuberosus;* Os - *Oryza sativa;* Pm - *Pseudotsuga menziesii;* Pl – *Picea glauca;* Ptri – *Populus trichocarpa;* Pt - *Populus tremula;* Sm - *Salvia miltiorrhiza;* St - *Solanum tuberosum;* Vv - *Vitis vinifera;* Zm - *Zea mays*.

The C/D box gene 119b (about 100 bp) was present in 19 species, and 22 homologous sequences were aligned (Figure 3.12). Two different homologues were found for *Gossypium arboceum* (Ga119b) *Lactua virosa* (Lv119b) and *Vitis vinifera* (Vv119b). Two putative primer sites were identified, one containing the antisense element (dotted line) and the box D' (15 bp, consensus: 5' GCACTGGGCTCTGAG) and the other near box D (20 bp, consensus: 5' AWTGCAGCCTAGAAAGCTAT) (Figure 3.12). An alternative possible primer site (21 bp, consensus: AGATGATGADTDTGCACTGGG), a

5’ end extension of the first 8 bases of the box D’ containing site was designed via an earlier alignment of a subset of sequences available.

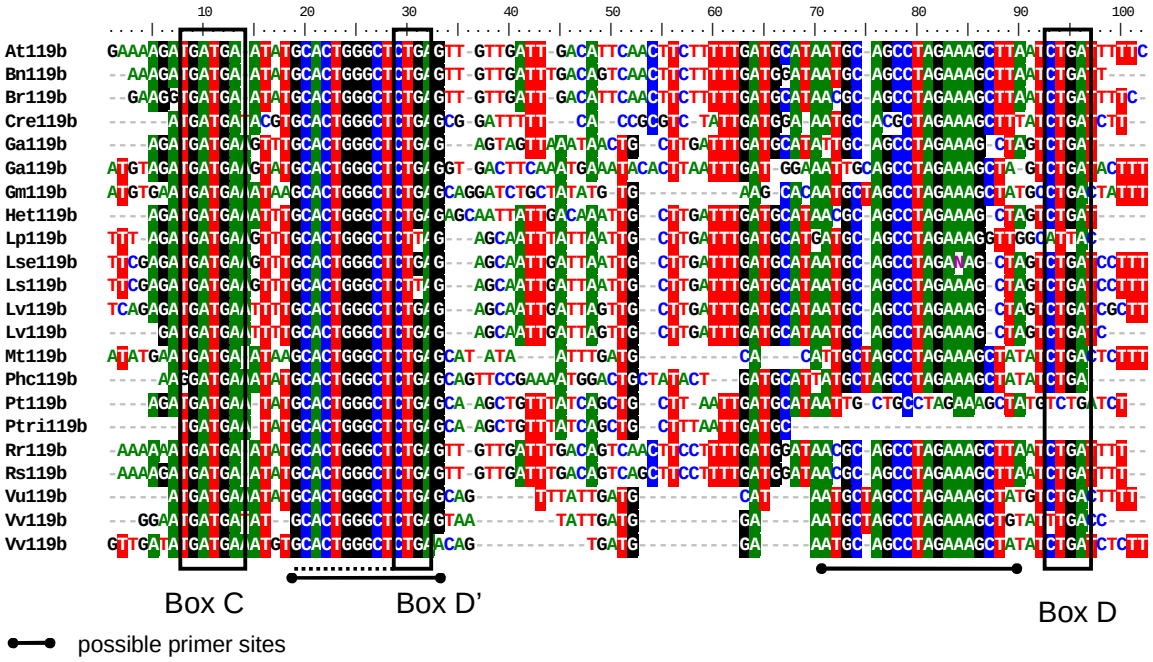


Figure 3.12: Box C/D gene 119b sequence alignment of twenty-two comprising eighteen species. Dotted line – antisense element sequence. Conserved positions are shaded. At – *Arabidopsis thaliana*; Bn - *Brassica napus;* Br – *Brassica rapa;* Ga - *Gossypium arboceum;* Het - *Helianthus tuberosus;* Lp – *Lactua perennis;* Lse - *Lactua serriola;* Lv – *Lactua virosa;* Rr - *Raphanus raphanistrum;* Mt - *Medicargo trunculata;* Phc - *Phaseolus coccineus;* Pt - *Populus tremula;* Ptri – *Populus trichocarpa;* Rs - *Raphanus sativa* ; Vu - *Vigna ungurculata;* Vv - *Vitis vinifera.*

3.3.1.4 SnoR37-SnoR22-SnoR23-SnoR80 gene cluster (cluster D)

Four snoRNA genes are present in the *A. thaliana* cluster (snoR37-snoR22-snoR23-snoR80) (Figure 3.2D). Although this gene cluster is present in three copies in *A. thaliana*, BLAST searches were conducted using only one of these copies, snoR37-1 and snoR80-1, respectively. Twenty-six box C/D snoR37 sequences (about 110 bp in length) found in twenty-one species were aligned and two putative primer sites were discovered (Figure 3.13). The first primer site (20 bp, consensus: 5’ GTGGACTAGAGTTTCHGATC) includes the antisense element and the box D’,

whereas the other (18 bp, consensus: 5’ AACCCTTGGCTGTCTGAG) contains the box D and the adjacent 13 bp upstream. Additional to the three *A. thaliana* homologues (AtsnoR37-1 – 3), the alignment contains two different homologues of *Euphorbia esula* and three of *Glycine max* (Figure 3.13).

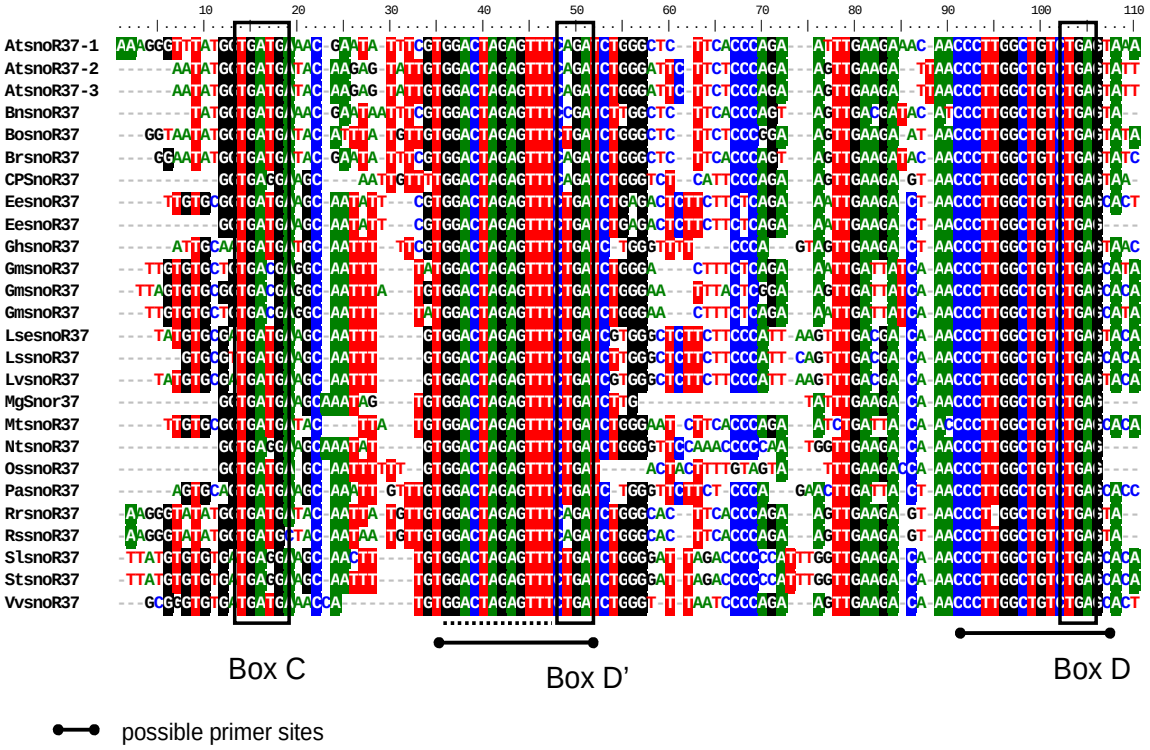


Figure 3.13: Alignment of twenty-six Box C/D gene snoR37 homologues obtained from twenty-one species. Conserved positions are shaded. Dotted line – antisense element sequence. At – *Arabidopsis thaliana*; Bn - *Brassica napus;* Bo - *Brassica oleraceae;* Br – *Brassica rapa* subsp. *pekinensis;* CP - *Carica papaya;* Ee - *Euphorbia esula;* Gh - *Gossypium hirsutum;* Gm – *Glycine max;* Lse - *Lactuca serriola;* Ls - *Lactua sativa;* Lv – *Lactua virosa;* Mg - *Mimulus gutatus;* Mt - *Medicargo trunculata;* Nt - *Nicotianum tabacum;* Os - *Oryza sativa;* Pa - *Populus alba;* Rr - *Raphanus raphanistrum* subsp. *maritimus;* Rs - *Raphanus sativa;* Sl - *Solanum lycopersicum;* St - *Solanum tuberosum*; Vv - *Vitis vinifera.*

Two putative primer sites were found using the alignment of 24 homologous snoR80 sequences (about 160 bp in length) obtained from twenty different species (Figure 3.14). The first conserved sequence (18 bp, consensus: 5’ TTACCAATTCTGRRGGAT) is

located at the beginning of the gene and the second (19 bp, consensus: 5' TTTGATCYTGAAABGCCMC) close to the box ACA at the end of the gene. Each sequence aligned belongs to a different species, except the three homologues present in *A. thaliana* and three different sequences of *Medicago trunculata* (Figure 3.14).

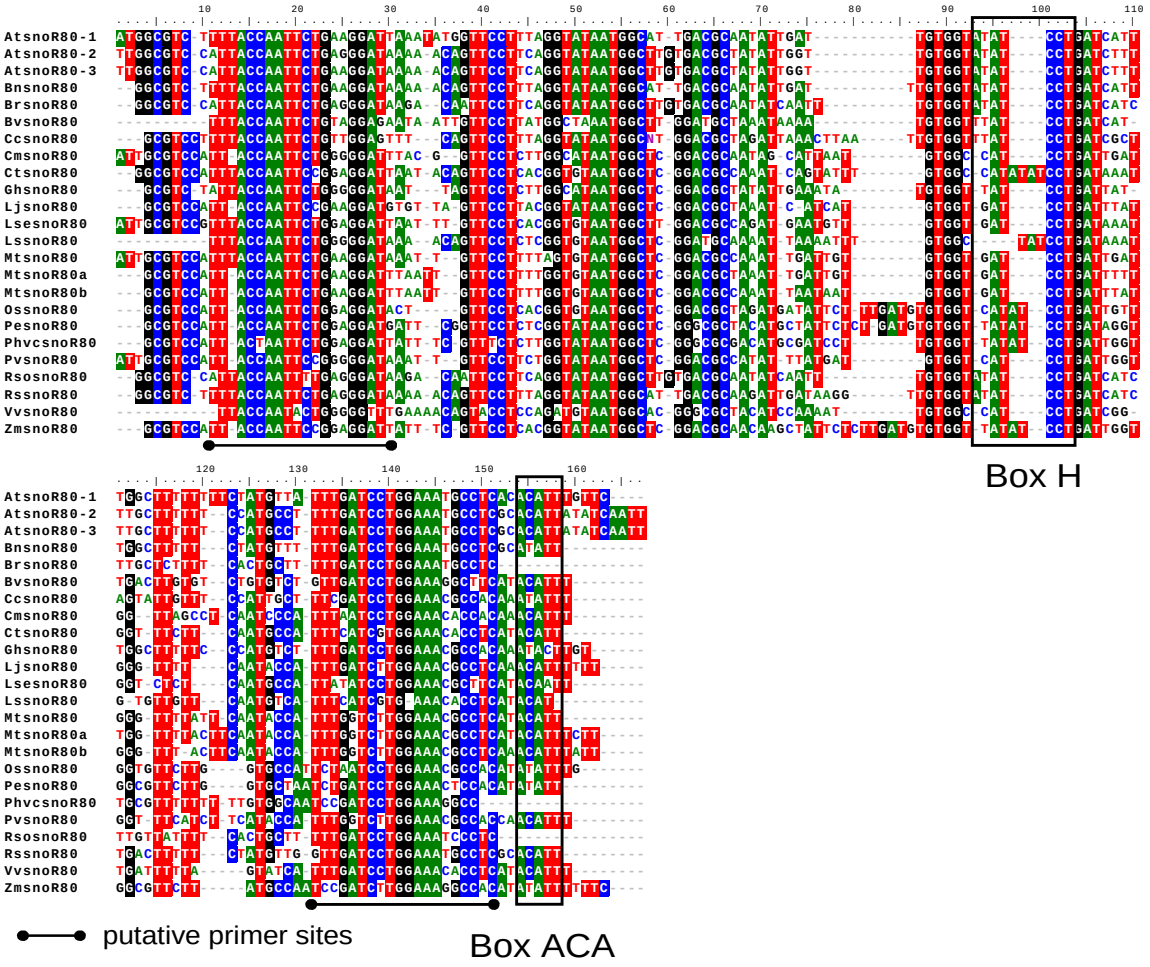


Figure 3.14: Alignment of twenty-four Box C/D gene snoR80 homologues obtained from twenty species. Conserved positions are shaded. At – *Arabidopsis thaliana*; Bn - *Brassica napus;* Br – *Brassica rapa* subsp. *pekinensis;* Bv - *Beta vulgaris* ; Cc - *Cistus creticus subsp.* creticus*;* Cm - *Cucumis melo* subsp. *melo;* Ct - *Carthamus tinctorius;* Gh - *Gossypium hirsutum;* Lse - *Lactuca serriola;* Ls - *Lactua sativa;* Mt - *Medicago trunculata;* Os - *Oryza sativa;* Pe - *Phyllostachys edulis;* Phv – *Phaseolus vulgaris;* Pv – *Panicum virgatum;* Rr - *Raphanus raphanistrum* var. *oleiformis;* Rr - *Raphanus raphanistrum;* Vv - *Vitis vinifera;* Zm - *Zea mays.*

3.3.1.5 snoR114-snoR115-snoR85 gene cluster (cluster E)

Three different snoRNA genes are found in the *A. thaliana* cluster snoR114-snoR115-snoR85 (Figure 3.2E). Nineteen homologues of the box C/D box snoR114 gene (about 100 bp in length) were aligned and one possible primer site (17 bp, consensus: 5' TTGTCCGTACCATCTGA), including box D' as well as a putative antisense element, was identified. Two homologous copies of this gene were found within the same EST of *Helianthus annuus* (HeasnoR114a and 114b) (Figure 3.15).

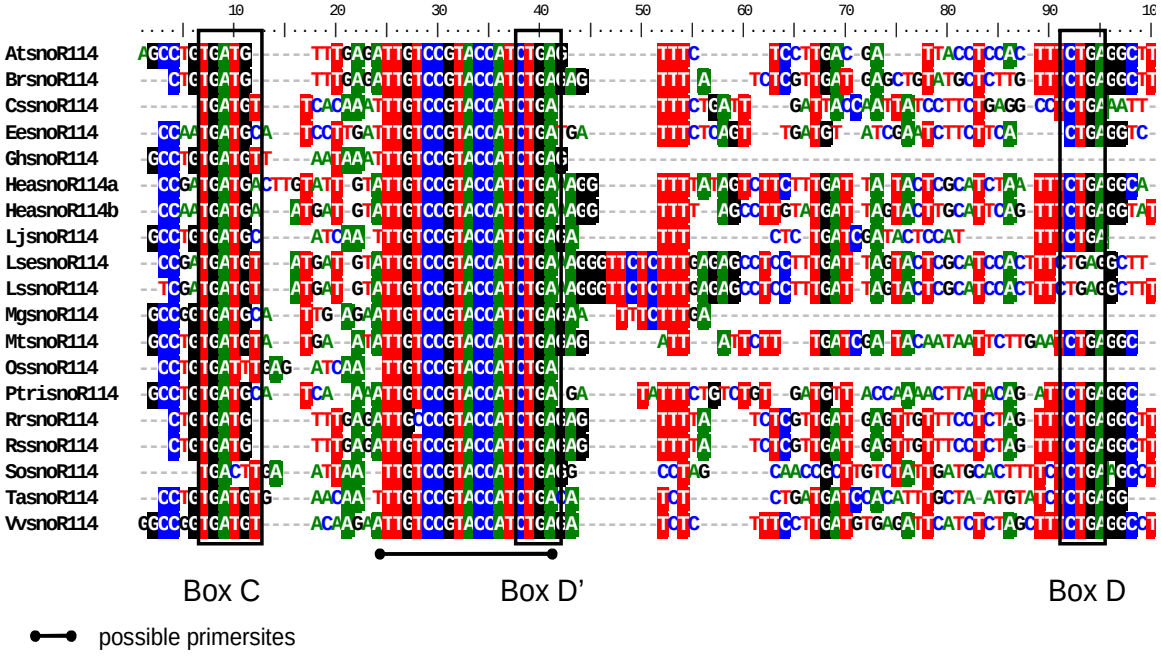


Figure 3.15: Alignment of nineteen homologous box C/D snoR114 sequences found in eighteen species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa* subsp. *pekinensis;* Cs - *Citrus sinensis;* Ee - *Euphorbia esula;* Gh - *Gossypium hirsutum;* Hea - *Helianthus annuus;* Lj - *Lotus japonicus;* Ls - *Lactua serriola;* Ls _ *Lactuca sativa;* Mg - *Mimulus gutatus;* Mt - *Medicago trunculata;* Os - *Oryza sativa;* Ptri – *Populus trichocarpa;* Rr - *Raphanus raphanistrum* subsp. *landra;* Rs - *Raphanus sativa;* Sa – *Saccharum officinarum;* Ta - *Triticum aestioum;* Vv - *Vitis vinifera.*

Ten box C/D snoR115 gene homologues (about 90 bp in length), found in 10 different species, were aligned and one possible primer site (21 bp:5' TACCATCAAACTTTGAGAGST) was identified containing the box D' and a putative antisense element (Figure 3.16).

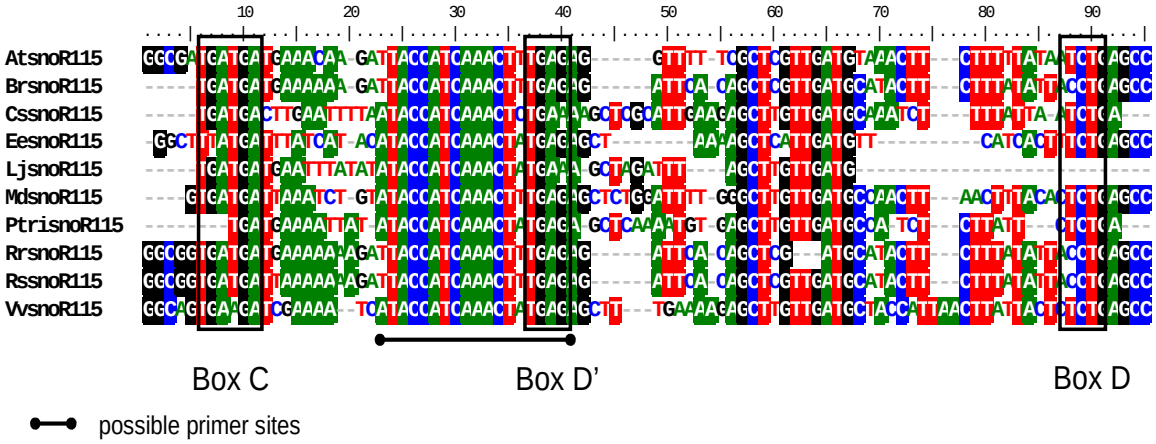


Figure 3.16: Alignment of ten box C/D snoR115 gene homologues found in ten species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa;* Cs - *Citrus sinensis;* Ee - *Euphorbia esula;* Lj - *Lotus japonicus;* Md - *Malus domestica;* Ptri – *Populus trichocarpa;* Rr - *Raphanus raphanistrum* subsp. *landra;* Rs - *Raphanus sativa* Vv - *Vitis vinifera.*

Twelve homologous sequences of the box H/ACA snoR85 gene were obtained from 11 species (two copies - AtsnoR85a and b were present in *Arabidopsis thaliana*). The alignment revealed one possible primer site (23 bp: 5' AAGGCAAYAAATTAGAGTCTCTG) at the beginning of the gene (Figure 3.17).

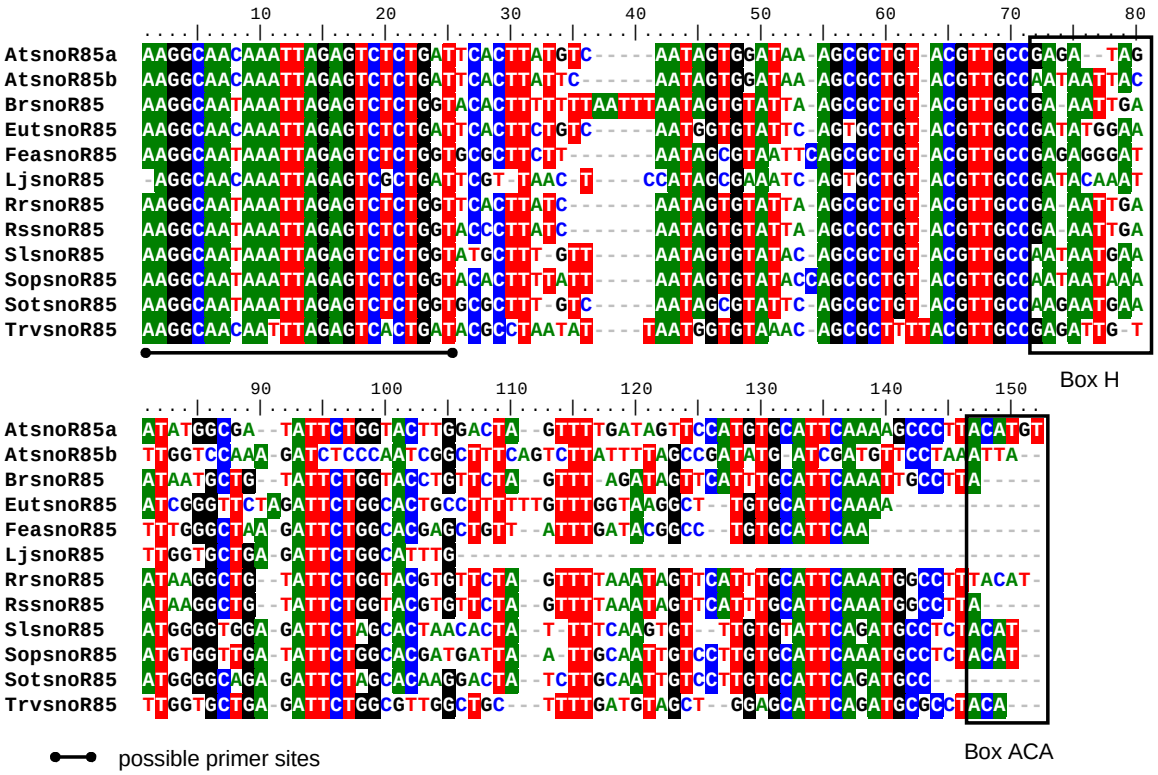


Figure 3.17: Alignment of twelve box H/ACA snoR85 gene homologues found in eleven species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa* subsp. *pekinensis;* Eut - *Euphorbia tiracalli;* Fea - *Festuca arundinaceae;* Lj - *Lotus japonicus;* Rr - *Raphanus raphanistrum;* Rs - *Raphanus sativa;* Sl - *Solanum lycopersicum;* Sop - *Solanum pennellii;* Sot - *Solanum tuberosum;* Trv - *Triphysaria versicolor.*

The alignment (Figure 3.18) shows a highly conserved region, chosen for a possible primer site, at the beginning of the gene. Besides the above sequences (Figure 3.17) the BLAST search also identified many more sequences homologous to the first 80 to 100 bp of snoR85. Like *A. thaliana*, it appears that most of the species contain two isoforms of the snoR85 gene related by sequence homology in the 5' half of the gene. In this analysis, only one copy was identified from *Citrullus lanatus*, *Cucumis melo* subsp. *agrestis*, *Cyamopsis tetragonoloba*, *Festuca arundinacea*, *Solanum lycopersicum*, *Solanum tuberosum* and three copies in *Ipomoea nil*. These additional sequences were placed in the alignment which now consists of 33 sequences (100 bp in length) from 19 different

species (Figure 3.18). The extended alignment (Figure 3.18) confirms the choice of the putative primer site.

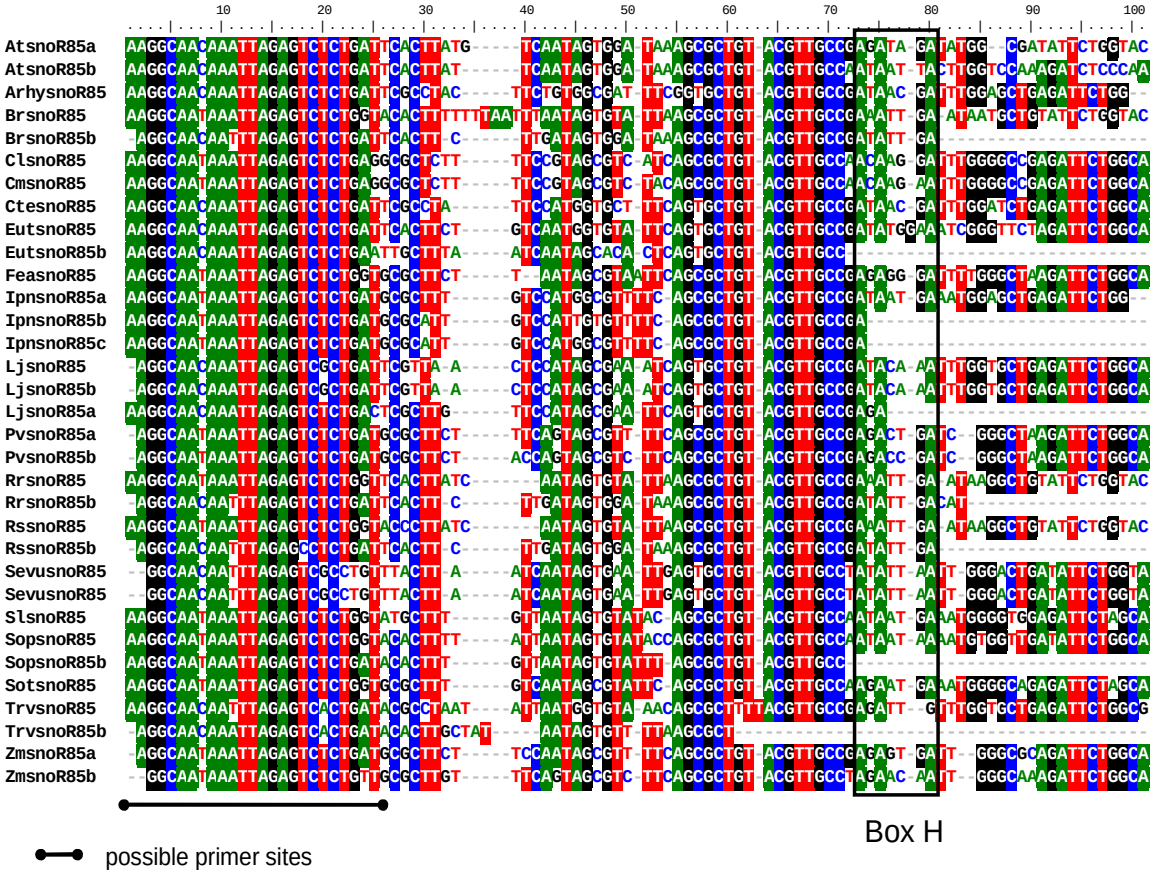


Figure 3.18: Extended alignment of the first 100 bp of thirty-three box H/ACA snoR85 gene homologues found in 19 species. Conserved positions are shaded. At – *Arabidopsis thaliana;* Br – *Brassica rapa* subsp. *pekinensis;* Cl - *Citrullus lanatus;* Cm - *Cucumis melo subsp. agrestis;* Cte - *Cyamopsis tetragonoloba;* Eut - *Euphorbia tiracalli;* Fea - *Festuca arundinaceae;* Ip - *Ipomoea nil;* Lj - *Lotus japonicus;* Pv - *Panicum virgatum;* Rr - *Raphanus raphanistrum;* Rs - *Raphanus sativa;* Sev - *Senecio vulgaris* subsp. *vulgaris*Sl - *Solanum lycopersicum;* Sop - *Solanum pennellii;* Sot - *Solanum tuberosum;* Trv - *Triphysaria versicolor.*

3.3.2 Gene order conservation in gene clusters

Gene clusters were examined for "*Arabidopsis* like" organisation of snoRNA genes in other species by examining the single gene BLAST sequences obtained for putative and expected neighbouring genes and/or by performing complete cluster sequence BLAST searches. Many of the sequence hits were to ESTs and represented transcripts from the gene clusters. While these contained one or more snoRNA gene sequences, it is likely that many representing gene organisations presented in Figure 3.19 may be incomplete. All five gene clusters tested could be found, at least partially, in various species (Figure 3.19). While the two gene clusters (or parts of them) 424-502-snoR95 (cluster A; Figure 3.19A) and snoACA1-snoR68-319-122-118a-118b (cluster B; Figure 3.19B) were only found in three and two other species, respectively, the three clusters (or parts of them) 382-snoR66-119b (cluster C; Figure 3.19C), snoR37-snoR22-snoR23-snoR80 (cluster D; Figure 3.19D) and snoR114-snoR115-snoR85a-snoR85b (cluster E; Figure 3.19E) were present in many plant species. Some gene clusters, however, showed a different gene organisation in several species, with parts of the gene clusters missing as well as genes being duplicated and/or rearranged. Cluster A (Figure 3.19A) containing snoRNA genes 424, 502 and snoR95 were found in only four species but the gene order appears to be conserved. Cluster B (Figure 3.19B) was present in two other species showing the same gene order. Cluster C (Figure 3.19C) consists of three genes, two of them (snoR66 and 119b) are found in many species and their gene order is highly conserved. Cluster D (Figure 3.19D), already present in three copies in *A. thaliana*, shows some level of conservation. The gene snoR37 examined was found upstream of snoR22, while snoR80 was downstream of snoR23 in most species. Cluster E (Figure 3.19E), containing snoR114, snoR115 and two copies of snoR85, showed some degree of conservation. Although snoR115 is mostly found upstream of snoR115, it is sometimes missing. The same is true for snoR115 and snoR85. In *Saccharum officinarum* the intergenic sequence between snoR114 and snoR115 (orange line, Figure 3.19E) is highly similar to the small ribonucleoprotein F in *Zea mays*.

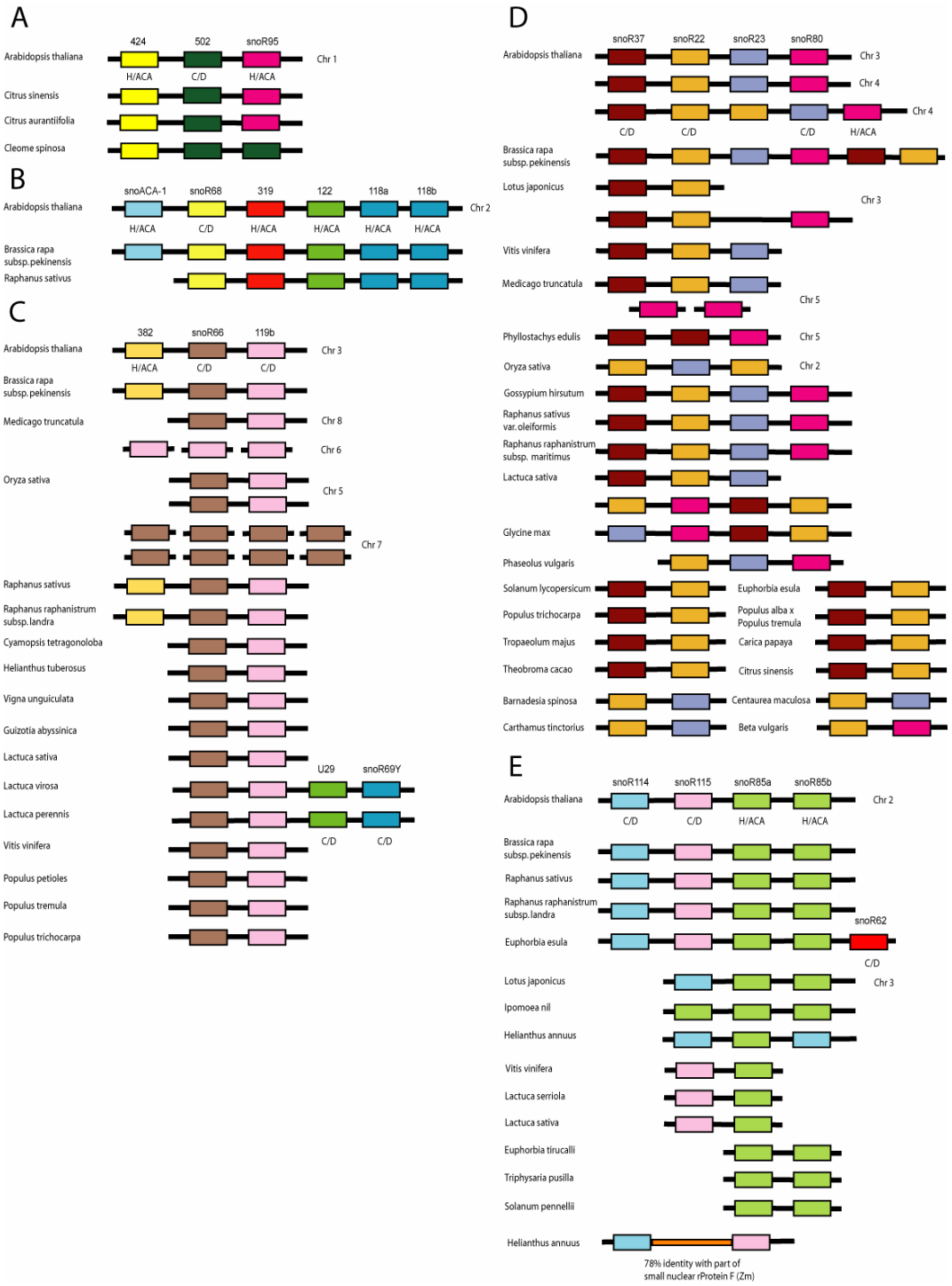


Figure 3.19: Gene order conservation in the gene clusters examined (A-E). Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes, the type of the snoRNA gene (i.e. box C/D and box H/ACA genes) below the boxes and the chromosome(s) where they are located to the right of a cluster.

Arabidopsis thaliana homologues of cluster A were found in *Citrus sinensis* and *Citrus aurantiifolia*. In *Cleome spinosa*, however, gene 502 was duplicated within the same cluster, while gene snoR95 could not be found (Figure 3.19A). This sequence should be viewed with caution because of the "TestAgain" tag in its name.

The six-gene cluster B was found in *A. thaliana*, *Brassica rapa* subsp. *pekinensis* and *Raphanus sativus*. In the latter species, however, gene snoACA-1 was missing (Figure 3.19B).

Cluster C (Figure 3.19C) consists of three genes in *A. thaliana*, *Brassica rapa* subsp. *pekinensis* and both *Raphanus* species, but gene 382 is missing in all other species. In *Oryza sativa* this snoR66-119b cluster was duplicated within the same chromosome and the snoR66 gene is present in 8 copies on chromosome 7 (Figure 3.19C). In addition to the snoR66-319b gene cluster on chromosome 8 in *Medicago trunculata*, three copies of gene 319b could be found on chromosome 6. In *Lactucta virosa* and *Lactuca perennis*, but not in *Lactuca sativa*, the U29 and snoR69Y genes (found in a gene cluster on a different chromosome in *A. thaliana* although in a different gene order) are attached to snoR66 and 119b (Figure 3.19C).

Homologues of the complete cluster D (Figure 3.19D), present in 3 copies in *A. thaliana*, were found *in Brassica rapa* subsp. *pekinensis*, *Gossypium hirsutum* and two species of *Raphanus* and parts of this cluster were identified in *Vitis vinifera*, *Medicago trunculata*, *Lactuca sativa* (snoR37-snoR22-snoR23), *Phaseolus vulgaris* (snoR22-snoR23-snoR80), *Lotus japonicus*, *Solanum lycopersicum*, *Populus trichocarpa*, *Populus alba* x *Populus tremula*, *Tropaeolum majus*, *Theobroma cacao*, *Carica papaya*, *Citrus sinensis* (snoR37-snoR22), *Oryza sativa*, *Barnadesia spinosa*, *Carthamus tinctorius* and *Centaurea maculosa* (snoR22-snoR23) (Figure 3.19D). Furthermore, in *Brassica rapa* subsp. *pekinensis* an additional snoR37-snoR22 was duplicated next to the snoR80 gene. Two copies of the snoR80 gene, although not adjacent to cluster D and each other, were present on chromosome 5 in *Medicago trunculata*. In addition to the snoR37-snoR22 gene cluster part on chromosome 3 in *Lotus japonicus* another cluster D missing snoR23 was identified. The cluster D found in *Phyllostachys edulis* has no snoR22 and snoR23, but does have two copies of snoR37, while the cluster found in *Beta vulgaris* consists of only snoR22 and snoR80. While all cluster D genes are present, although in different

order, in *Glycine max*, snoR23 was substituted by a second snoR22 gene in one of the two cluster D copies in *Lactuca sativa* (Figure 3.19D).

Complete cluster E (Figure 3.19E) homologues of *A. thaliana* were found in *Brassica rapa* subsp. *pekinensis*, *Raphanus sativus, Raphanus raphanistrum* subsp. *landra* and *Euphorbia esula* and parts of it in 10 other species. While *Lotus japonicus* contains snoR114 and the two snoR85 genes, only snoR115 and one snoR85 gene are present in *Vitis vinifera, Lactuca sativa* and *Lactuca serriola*. *Euphorbia tirucalli*, *Triphysaria pusilla* and *Solanum pennelli* contain a cluster consisting of two snoR85, and *Ipomoea nil* contains one of three snoR85 genes. *Helianthus annuus* harbours one snoR85 gene between two copies of snoR114, and a snoR62 gene was found downstream to cluster E in *Euphorbia esula*. Surprisingly, in *Saccharum officinarum* the sequence separating two snoR114 genes (orange line, Figure 3.19E) is highly similar (78 %) to parts of the small nuclear ribonucleoprotein F gene found in *Zea mays*. Furthermore, a BLAST search of this intergenic region revealed that a sequence of about 170 bp in length is present in a wide range of species but in multiple copies on almost every chromosome in rice as well (Figure 3.19E).

3.3.3 Virtual amplification of primer combinations using reverse ePCR

When taking the gene order of the snoRNA genes into consideration, pairs of primers were chosen consisting of one forward and one reverse primer. Reverse primers were obtained by generating reverse complements of the putative primer sequences (Table 3.1) using BioEdit version 7.0.9.0. Although the sequence of a reverse complement is different, its characteristics (length, GC content and TM) remain the same. The primers on the edges of a gene cluster (the upstream-most and downstream-most primers) were only needed in the forward and backward directions. The primers located within the cluster were designed in both directions (Table 3.2).

Table 3.2: Forward and backward primers designed for each gene cluster. Sequences in italics were chosen for further experiments.

primer name	sequence	primer name	sequence
424-502-snoR95 cluster (cluster A)		snoR66-1aR	TCAGGTCTCAAAGATWWCATGCCATC
424F	ATAGCCCCTTGCWWCTT	snoR66-1bF	TGATGGCATGAAATCTTTG
502F	CTTCAAAGTTCTCTGA	snoR66-1bR	CAAAGATTTCATGCCATCA
502R	TCAGAGAACTTTGAAG	119b-1aF	AGATGATGADTDTGCACTGGG
snoR95R	GCACCATGCTCGTRTAG	119b-1aR	*CCCAGTGCAHAHTCATCATCT*
snoACA1-snoR68-319-122-118a-118b cluster (cluster B)		119b-1bF	GCACTGGGCTCTGAG
snoR68F	TGGTTCGTATTCVCTGAGCA	119b-1bR	CTCAGAGCCCAGTGC
319F	CCAAGTTTRCCTTCGDAWAT	119b-2R	*CTTTCTAGGCTGCAWTATGCATC*
319R	ATWTHCGAAGGYAAACTTGG	**snoR37-snoR22-snoR23-snoR80 cluster (cluster D)**	
122-1F	GCGAAGGDCCCAGCAGRG	snoR37-1F	GTGGACTAGAGTTTCHGATC
122-1R	CYCTGCTGGGHCCTTCGC	snoR37-2F	AACCCTTGGCTGTCTGAG
122-2F	TGAGDCYTCTCTAACAAT	snoR37-2R	CTCAGACAGCCAAGGGTT
122-2R	ATTGTTAGAGARGHCTCA	snoR80-1F	TTACCAATTCTGRRGGAT
118-1F	GTGTGTATCGGCKTWGTGC	snoR80-1R	ATCCYYCAGAATTGGTAA
118-1R	GCACWAMGCCGATACACAC	snoR80-2R	GKGGCVTTTCARGATCAAA
118-2F	AGRTGGGCAGTTGTGHTTCA	**snoR114-snoR115-snoR85a-snoR85b cluster (cluster E)**	
118-2R	TGAADCACAACTGCCCAYCT	snoR114F	*TTGTCCGTACCATCTGA*
118-3R	TGTRGGGAARATGATTGTTGA	snoR115F	*TACCATCAAACTTTGAGAGST*
382-snoR66-119b cluster (cluster C)		snoR115R	*ASCTCTCAAAGTTTGATGGTA*
382F	GCARGGGCGYTGAGTCGCTT	snoR85F	AAGGCAAYAAATTAGAGTCTCTG
snoR66-1aF	*GATGGCATGWWATCTTTGAGACCTGA*	snoR85R	*CAGAGACTCTAATTTRTTGCCTT*

To determine the number and lengths of possible amplification products, primer pairs were tested by virtual PCR against the *A. thaliana* and *Oryza sativa* genome reference and transcriptome snapshot databases (Table 3.3). No matches were obtained for the searches against the transcriptome snapshot databases, but various sequences were found in the genome databases and amplification success varied between certain gene clusters using the primer sequences shown in Table 3.2. For example in cluster B for which 21 primer pairs were available only four combinations amplified a virtual product. In contrast every primer pair for cluster E resulted in a fragment (not shown). However, many of these primers contain wobble bases which are likely to be the reason for non-amplification. For example, there was no amplification whenever a 319 primer (cluster B) was involved due to three wobble bases in its sequence (N>2). These wobble bases were substituted with *A. thaliana* corresponding bases (refined *A. thaliana* primers). Additionally, for the primer combinations designed for cluster A, for every possible

primer sequence (14 different sequence combinations – completely refined sequences) pair a reverse ePCR was performed.

Arabidopsis thaliana primer sequences revealed two primers (319-1b and 118-3) which did not give any amplification due to mistakes (missed out bases) made during the primer design (not shown). After the redesign of these primers expected fragments were obtained. Most of the other primer combinations led to the amplification of expected products but sometimes unexpected fragments were obtained as well (see Table 3.3) in *A. thaliana*. Many of the fragments obtained could be linked to gene references. Most of these references did not contain further information but a few identified the products correctly and, interestingly, four of them referred to protein coding genes (see below). One unexpected fragment was obtained for cluster A and cluster C (in italics, Table 3). The 424F/502R primer pair showed an unexpected fragment of 207 bp from chromosome 2, referred to as AT2G28105 – a hypothetical protein with no further specification. The 502F/snoR95R primer combination amplified a fragment of 629 bp from chromosome 1, a part of ATSS3 – a starch synthase/transferase. The snoR66-1bF/119b-1bR did not only amplify an expected fragment of 170 bp on chromosome 3, but also an unexpected one of 947 bp on chromosome 5, referred to as AT5G28495 – a transposable element gene which belongs to the gypsy-like retrotransposon family, which also matched a reverse transcriptase in *Sorghum bicolor*.

Performing a reverse ePCR against the *Oryza sativa* genome revealed that most primer combinations showing virtual amplification in *A. thaliana* were absent in *Oryza sativa* (Table 3.3). For instance, no product was obtained for cluster A and only five, three, one and two primer pairs were successfully amplified for cluster B, C, D and E, respectively (Table 3.3). Only in two cases, snoR66F/119b-1R and snoR37-2F/snoR80-1R, did the number of products in rice exceed that expected in *A. thaliana* and three and four fragments, respectively, were obtained for these primer pairs. These fragments did not match any gene reference with the exception of the fragment obtained with the snoR66-2F/119b-2R primer pair. This fragment is 1318 bp in length, located on chromosome 11 and referred to as Os11g0157000 – a hypothetical catalytic region domain containing protein.

Table 3.3 Table 3: Reverse e-PCR for primer combinations of various snoRNA gene cluster. Chr. no. (+/-) – Chromosome number and strand (+/-) from which products were virtually amplified.

Primer pair	*Arabidopsis thaliana*				*Oryza sativa*			
	Amplification	Product size (bp)	Chromosome	Fragment information	Amplification	Chromosome (+/-)	Product size (bp)	Fragment information
424-502-snoR95 cluster (cluster A)								
424F/502R	√	286	1 (-)	Expected				
		207	*2 (+)*	*AT2G28105*[a]				
424F/snoR95R	√	459	1(-)	Expected				
502F/snoR95R	√	189	1 (-)	Expected				
		629	*1 (-)*	*ATSS3*[b]				
snoACA1-snoR68-319-122-118a-118b cluster (cluster B)								
snoR68F/319R	√	118	2 (+)	Expected	√	5(+)	177	-
snoR68F/122-1R	√	264	2 (+)	Expected	√	6 (+)	93	-
snoR68F/122-2R	√	302	2 (+)	Expected	√	5 (+)	359	-
snoR68F/118-1R	√	455	2 (+)	Expected				
		750	2 (+)	Expected				
snoR68F/118-2R	√	525	2 (+)	Expected				
		821	2 (+)	Expected				
snoR68F/118-3R	√	562	2 (+)	Expected				
		859	2 (+)	Expected				
319F/122-1R	√	165	2 (+)	Expected				
319F/122-2R	√	203	2 (+)	Expected	√	12 (-)	1289	-
319F/118-1R	√	356	2 (+)	Expected				
		651	2 (+)	Expected				
319F/118-2R	√	426	2 (+)	Expected				
		721	2 (+)	Expected				
319F/118-3R	√	463	2 (+)	Expected				
		760	2 (+)	Expected				
122-1F/122-2R	√	56	2 (+)	Expected				
122-1F/118-1R	√	209	2 (+)	Expected				
		504	2 (+)	Expected				
122-1F/118-2R	√	279	2 (+)	Expected				
		575	2 (+)	Expected				
122-1F/118-3R	√	316	2 (+)	Expected				
		613	2 (+)	Expected				
122-2F/118-1R	√	171	2 (+)	Expected				
		466	2 (+)	Expected				
122-2F/118-2R	√	241	2 (+)	Expected				
		537	2 (+)	Expected				
122-2F/118-3R	√	278	2 (+)	Expected				

		575	2 (+)	Expected				
118-1F/118-2R	√	89	2 (+)	Expected				
		385	2 (+)	Expected				
		90	2 (+)	Expected				
118-1F/118-3R	√	126	2 (+)	Expected	√	7 (-)	356	-
		423	2 (+)	Expected				
		128	2 (+)	Expected				
118-2F/118-3R	√	57	2 (+)	Expected				
		354	2 (+)	Expected				
		58	2 (+)	Expected				
382-snoR66-119b cluster (cluster C)								
382F/snoR66-1aR	√	139	3 (-)	Expected				
382F/snoR66-1bR	√	131	3 (-)	Expected				
382F/119b-1bR	√	282	3 (-)	Expected				
382F/119b-2R	√	332	3 (-)	Expected				
snoR66-1aF/119b-1bR	√	169	3 (-)	Expected	√	3 (-)	198	-
						5 (-)	221	-
						5 (-)	220	-
snoR66-1aF/119b-2R	√	219	3 (-)	Expected				
snoR66-1bF/119b-1bR	√	170	3 (-)	Expected	√	4 (-)	36	-
		947	*5 (-)*	*AT5G28495*[c]		4 (-)	361	-
snoR66-1bF/119b-2R	√	220	3 (-)	Expected		11 (-)	*1318*	*Os11g0157000*[d]
119b-1aF/119b-2R	√	79	3 (-)	Expected				
119b-1bF/119b-2R	√	65	3 (-)	Expected				
snoR37-snoR22-snoR23-snoR80 cluster (cluster D)								
snoR37-1F/snoR37-2R	√	67	3 (+)	Expected				
		67	4 (+)	Expected				
		67	4 (-)	Expected				
snoR37-1F/snoR80-1R	√	447	3 (+)	Expected				
		502	4 (+)	Expected				
		678	4 (-)	Expected				
snoR37-1F/snoR80-2R	√	551	3 (+)	Expected				
		604	4 (+)	Expected				
		780	4 (-)	Expected				
snoR37-2F/snoR80-1R	√	398	3 (+)	Expected	√	3 (+)	98	-
		453	4 (+)	Expected		3 (+)	264	-
		629	4 (-)	Expected		6 (+)	308	-
						6 (+)	131	-
snoR37-2F/snoR80-2R	√	502	3 (+)	Expected				
		555	4 (+)	Expected				
		731	4 (-)	Expected				
snoR80-1F/snoR80-2R	√	122	3 (+)	Expected				
		120	4 (+)	Expected				
		120	4 (-)	Expected				
snoR114-snoR115-snoR85a-snoR85b (cluster E)								
snoR114F/snoR115R	√	130	1 (-)	Expected				
snoR114F/snoR85R	√	438	1 (-)	Expected	√	8 (-)	213	-
		255	1 (-)	Expected				
snoR115F/snoR85R	√	329	1 (-)	Expected				
		146	1 (-)	Expected				

snoR85F/snoR85R	√	23	1 (-)	Expected	√	4 (-)	23	-
		23	1 (-)	Expected		8 (-)	23	-
		206	1 (-)	Expected				

[a] hypothetical protein

[b] starch synthase/transferase

[c] Transposable element gene, gypsy-like retrotransposon family, with an 8.8e-77 P-value to a reverse transcriptase found in Sorghum bicolour

[d] hypothetical protein containing an integrase, a catalytic region domain

Reverse ePCRs using complete refined primer pairs against the *A. thaliana* genome database resulted in 20 different fragments from all chromosomes for 424F/502R, three from chromosome 1 and 3 for 424F/snoR95 and two from chromosome 1 for 502F/snoR95 (not shown). The references to sequences obtained include sequences such as transposable element genes, tetrahydrofolylpolyglutamate synthase and F-box family protein. Furthermore, using the *Oryza sativa* genome database multiple hits were obtained. Even for both transcriptome snapshot databases fragments were obtained, which was not observed using both wobble containing and refined *Arabidopsis* primers.

3.4 Discussion

3.4.1 Blast searches using single *Arabidopsis thaliana* gene sequences

BLAST searches using whole gene sequences resulted in a low number of homologues for some snoRNA genes. Genes might actually be absent in some species, either because the gene was never present or was lost in time. If the gene was never present in a species it must have originated after the diversification of the particular species lineage and that of *A. thaliana*. In the case of gene loss, various mechanisms might have contributed. The gene function could have been lost due to mutations and selection leading to pseudogene production and subsequent loss of the gene. Also, unequal crossing over and gene conversion could be responsible for the disappearance of a gene (Barneche *et al.*, 2001; Qu *et al.*, 2001; Brown *et al.*, 2003a). Although the absence of a gene cannot be ruled out, it is more likely, at least for differences in the number of homologues found between

clusters that the lack of homologous sequences is due to incomplete EST libraries being available. For example, rarely transcribed genes might not be present in the available EST libraries (Bonaldo *et al.*, 1996).

The difference in number of homologues found between genes of the same cluster might be best explained by different gene organization. A gene with low homologue number might not be part of the same cluster in other species and may either be lost or present in a rarely transcribed region. Another possibility, although unlikely, is that homologous genes/gene clusters may have diverged to an extent which does not allow identification by BLAST analysis. The investigated gene clusters contain both box H/ACA and box C/D snoRNA genes. The former produce their antisense elements by forming secondary structure stem loops (e.g. Brown *et al.*, 2003a; Makarova & Kramerov, 2007) and, thus, lack a longer conserved primary sequence. Furthermore, their boxes are shorter and their consensus sequences highly degraded. Although these features make it quite difficult to identify these genes (Brown *et al.*, 2003a), H/ACA homologues were usually found in similar numbers to homologues of box C/D genes. The two genes, 382 and snoACA-1, for which only 2 and 4 homologues were found, might therefore be either absent or copied to rarely transcribed areas in other species. At least they are not part of the same cluster, found in *A. thaliana*, in some other species.

3.4.2 Conservation and differences in the organization of gene clusters

Both differences and conservation in gene order could be shown for every gene cluster and it appears that the order of certain genes within a cluster is normally highly conserved. For instance, snoR66 and 119b were found in the same order in many species whereas gene 382 is often missing. Furthermore, the snoR37 and snoR22 combination and the two copies of snoR85 are present in most species. In some species however, even these conservations of gene order could not be observed. Different gene organization might be caused by deletions, insertions, conversion of genes/part of gene. Genes and parts of gene clusters might be cis-copied leading to the expansion of a gene cluster or trans-copied to other regions in the genome where they might be established leading to

new paralogs (Brown *et al.*, 2001; Brown *et al.*, 2003a; Chen *et al.*, 2008; Schmitz *et al.*, 2008). For instance, in *Oryza sativa* snoR66 and 119b are present twice on chromosome 5 suggesting a duplication event on the same chromosome. Furthermore, snoR66 can be found in 8 copies on chromosome 7 which might be the result of one trans- followed by various cis-duplications. A similar case of possible trans- and cis-duplication events can be seen in *Medicago trunculata* where snoR66 and 119b can be found on chromosome 8, but three additional copies of 119b are present on chromosome 6. However, a highly conserved gene order is necessary for the amplification of snoRNA genes and gene clusters, especially for genes containing only one putative primer site,. Thus, the gene clusters examined, or at least parts of them, can be used due to their relatively high conserved gene order. Comparing snoRNA gene clusters between species, particularly more distantly related ones, might reveal different gene organisations. In such cases, it might be best to extract the homologous sequences, particularly genes, for phylogenetic investigation. For closely related species, the complete and chosen parts of cluster sequences, respectively, might be compared because cluster reorganisation does not appear very likely. However, gene and gene cluster duplication as well as cluster reorganisation, although not very likely to happen between closely related species, should always be taken into consideration when working with these genes. While orthologous snoRNA gene/gene cluster sequences might be used to discover the phylogenetic history of species and for DNA barcoding, the differences in the organisation of gene cluster as well as possible duplications and deletions might be highly useful for studying snoRNA gene/gene cluster evolution and the reorganisation and transposition process during the evolution of different plant lineages (Brown *et al.*, 2003a).

3.4.3 Virtual amplification of designed primer pairs

Fragments of all *A. thaliana* clusters obtained from different chromosomes in *Oryza sativa* were found using reverse ePCR suggesting either trans-duplication events and putative differences in cluster organisation or the amplification of unintentional products. As *O. sativa*, a monocot, and *A. thaliana*, a dicot, are phylogenetically very distant relatives, another very likely possibility might be that the sequences contain snoRNA

genes but differ to such an extent that no significant similarity was found. Differences in cluster organisation between these two species could be shown in many clusters (Chen *et al.*, 2003) and, thus, it would not be surprising to find some gene clusters examined to be duplicated, reorganized and dispersed in *O. sativa*. Some of the *O. sativa* gene cluster and genes, however, were found by BLAST searches and sequences of this species were integrated in some of the alignments showing a high degree of conservation.

The main goal of this study, however, is to discover and examine snoRNA genes and gene clusters useful for phylogeny and DNA barcoding. It is, therefore, highly desirable to design primers which amplify only one fragment. Thus, all possible primer combinations were virtually tested. For each gene cluster, at least one, and for most clusters more primer combinations were virtually amplified. The difference in amplification success between primers containing wobbles and refined *A.thaliana* primers can be explained by the number of wobble bases used for some primers. As wobble bases cause mismatches by default, some primers might exceed the number of allowed mismatches leading to non-amplification. Although all primers were designed using *A. thaliana* homologues, some combinations resulted in amplification in *O. sativa* as well, suggesting universality of some primers. Unsurprisingly, the amplification success was less in *O. sativa* than in *A. thaliana* because wobbles were only substituted by *A. thaliana* corresponding bases, except for cluster A. Using different substitutes might increase the number of primer pairs leading to amplification. Using completely refined primer pairs for cluster A resulted in amplification of each primer pair. Unfortunately, these primer combinations led to the production of many unexpected fragments in both *A. thaliana* and *O. sativa*. Thus, using wobble containing primers of cluster A would amplify multiple fragments and, therefore, these primers should not be used for experiments. The snoR66-1bF/119b-1bR primer combination of cluster C, although not complete refined, amplified an unexpected fragment as well and should be discarded.

The reverse ePCR searches were conducted with two gaps and two mismatches allowed and, thus, some of the fragments obtained might not be amplified using real PCR. Therefore, snoR66-1F, 119b-1aF, 119b-1aR and 119b-2R from cluster C, snoR80

from cluster D and snoR114F, snoR115F, snoR115R and snoR85R from cluster E (sequences in italics, Table 3.2) were chosen and used in further experiments.

3.5 Conclusions

Five gene clusters containing 18 snoRNA genes, some of which are present in multiple copies, were investigated for conserved sequences suitable for primer binding sites. Using some freely available genomic tools these conserved regions were identified and 37 primers (including backward primers) were designed, characterized and virtually tested. In the end, only 9 primers, taken from three gene clusters, were selected for further use in experiments.

Chapter 4: SnoRNA gene/gene cluster length polymorphism (SRLP): A novel universal marker system for phylogenetic studies in *Senecio*

4.1 Introduction

Universal markers for phylogenetic analysis should be present in a wide range of species, have highly conserved regions for primer annealing, and be variable (Alvarez & Wendel, 2003; Chapman *et al.*, 2007). Depending on the particular application of these markers, different degrees of variation are needed (Small *et al.*, 1998; Small *et al.*, 2004). For example, at higher taxonomic scales, comparing genomic regions exhibiting relatively low variation (e.g. gene regions) is sufficient (Soltis *et al.*, 2000; Chase, 2001), whereas more variable regions (e.g. noncoding regions) are required for the analysis of more closely related taxa (Matthee *et al.*, 2007; Shaw *et al.*, 2007), while markers used in DNA barcoding should be sufficiently variable for species identification (Chase *et al.*, 2007; Hollingsworth *et al.*, 2009b).

SnoRNA genes and gene clusters are potential universal molecular markers for phylogenetic analysis. They are found in all eukaryotes and have highly conserved regions (e.g. antisense elements and adjacent boxes) for primer annealing. Amplicons (i.e. regions between primer sites) should be highly variable because they consist of gene regions that do not code for proteins and/or intergenic regions (Brown *et al.*, 2003a; Makarova & Kramerov, 2007).

DNA sequence variation is caused mainly through nucleotide substitution and/or insertions/deletions (indels) and the degree of variability of a region depends on their rate of occurrence (Britten *et al.*, 2003; Yamane *et al.*, 2006). In noncoding nuclear regions the rate for nucleotide substitutions is approximately 10 times higher than the rate for indels (Saitou & Ueda, 1994). However, while nucleotide polymorphisms can only be spotted by comparing DNA sequences, indels can also be detected by differences in fragment length (fragment lengths polymorphisms). Although the evolution of indels is still not completely understood (Kelchner, 2000) various models have been developed, e.g. the stepwise mutation model (Kimmel & Chakraborty, 1996; Fu & Chakraborty,

1998; Balloux & Goudet, 2002), the TKF91 (Thorne *et al.*, 1991) and TKF2 model (Thorne *et al.*, 1992), and the long indel model (Miklos *et al.*, 2004). Fragment length differences are thought to increase with genetic distance, such that closely related species are assumed to show more similar fragments than more distantly related taxa. Furthermore, assuming that different copies of genes or gene clusters, as well as different alleles of genes, produce fragments of different length, the minimum number of gene/gene cluster copies present in a species can be estimated by the number of different fragments amplified. Additionally, fragments that differ considerably in length might indicate different gene/gene cluster copies rather than different alleles of a particular gene. It should be noted that gene clusters containing homologous genes will generate fragment length differences from the same cluster and, therefore, gene copy estimations using length differences should be done with caution. Single copy regions produce a maximum of two different fragments (two alleles) in diploid species and, thus, it is possible to identify different alleles exhibiting codominance. However, because the alleles of a certain gene/gene cluster cannot be clearly identified without sequencing, fragment length variation between and within species can only be examined by treating the fragment profiles (i.e. fragment pattern of an individual) as dominant markers.

Two different strategies have been established for analysing population and phylogenetic structure using dominant markers (Hollingsworth & Ennos, 2004). First, to calculate a reliable distance matrix for assessing relationships among samples by means of a Neighbour Joining (NJ) tree or Principal Coordinate Analysis (PCO), a low number of samples should be examined across a large number of loci. Unresolved star-like trees where most samples are intermingled might result from data sets for which only a few fragments were amplified (e.g. single or low copy regions). Furthermore, in trees based on only a low number of fragments, a single band can have a high impact on topology. Second, a substantial number of individuals per group (e.g. population or species) should be used in a population based approach. Here, fragment frequencies are used to estimate genetic variation which can be partitioned within and among group components to calculate Φst (Excoffier *et al.*, 1992).

To investigate variation of snoRNA genes and gene clusters between and within *Senecio* species, these regions were amplified using the universal primers designed as

described in Chapter 3. Amplified fragment profiles were then examined for fragment length polymorphisms. In an initial investigation, described in this chapter, a small number of DNA samples comprising a wide variety of different *Senecio* species was explored by radioactive labelled fragment analysis. Fluorescence labelling was then used to examine some species in more detail using a greater number of individuals per species. Whereas in the initial screen only individual fragment analysis approaches were used, for the more detailed examinations, datasets were also analysed by population based approaches.

Shared fragments between parents and hybrids are expected to be present for at least some gene clusters. SnoRNA gene clusters are spread across the entire genome in *Arabidopsis thaliana* and this is also likely to be the case in *Senecio.* Consequently, the genomic contribution of parents to a hybrid species should be possible to estimate using snoRNA markers. Thus the work reported in this chapter was conducted to test primer pairs designed to amplify snoRNA genes and clusters, to explore fragment variation for several snoRNA gene clusters in and between various species of *Senecio*, and to determine how useful snoRNA markers are for detecting hybrids.

4.2 Material and Methods

4.2.1 Plant Material

Leaf material for DNA extraction was obtained from plants cultivated in the greenhouse. For an initial primer-trial the following 28 accessions (ac) of 13 different species/subspecies were examined: *S. aethnensis* (1 ac), *S. chrysanthemifolius* (1 ac), *S. squalidus* (2 ac), *S. massaicus* (2 ac), *S. vulgaris* ssp. *hibernicus* (1 ac), *S. vulgaris* ssp. *vulgaris* (1 ac), *S. glaucus* ssp. *coronopifolius* (3 ac), *S. cambrensis* (1 ac), *S. flavus* (3 ac), *S. teneriffae* (2 ac), *S. mohavensis* ssp. *mohavensis* (2 ac), *S. mohavensis* ssp. *breviflorus* (2 ac) and *S. madagascariensis* (7 ac) (see Table 2.1).

For a more detailed analysis of variation within and between species, 82 additional accessions of 8 different species were examined, together with 15 of the samples (1 – 15, 9 species) used in the initial primer-trial. Thus the numbers of accessions per species subjected to more detailed analysis were: *Senecio aethnensis* (11 ac), *S.*

chrysanthemifolius (12 ac), *S. squalidus* (29 ac), *S. vulgaris* (13 ac), *S. cambrensis* (12 ac), *S. madagascariensis* (9 ac), *S. teneriffae* (3 ac), *S. massaicus* (2 ac), *S. flavus* (3 ac), *S. glaucus* (1 ac) and *S. engleranus* (2 ac).

4.2.2 DNA-Extraction, PCR-amplification and fragment analysis

Total DNA was extracted from either frozen or fresh leaves. Leaf tissue was pulverized to a fine powder using liquid nitrogen and DNA was isolated using a modified 2 x CTAB (hexadecyltrimethyl ammonium bromide) extraction method (Doyle & Doyle, 1987). In the initial PCR analysis, 18 primer pairs derived from ten different gene clusters were used (for detailed primer information see Chapter 2, Table 2.2): U31F/snoR4R, U31F/U33R, U31F/U51R, snoR4F/U33R and U33F/U51R for cluster A (Figure 4.1A), U14-1/U14-2 and U14-3/U14-4 for cluster B (Figure 4.1B), U36aF/U38R for cluster C (Figure 4.1C), U49F/snoR2R and snoR2F/snoR77YR for cluster D (Figure 4.1D), snoR13F/U18R and U18F/U54R for cluster E (Figure 4.1E), U61F/snoR14R for cluster F (Figure 4.1F), snoR29F/SnoR30R for cluster G (Figure 4.1G), U80F/U80R for cluster H (Figure 4.1H), U15F/snoR7R for cluster I (Figure 4.1I) and snoR37F/Sno22R and snoR22F/SnoR23R for cluster J (Figure 4.1J).

In addition to the primer pairs used in the initial primer-trial study, ten extra primer combinations and two gene clusters were examined in the more detailed investigation that followed. These were: U49F/snoR77YR for cluster D (Figure 4.1D), snoR13F/U54R for cluster E (Figure 4.1E), snoR37F/snoR23R and snoR37F/snoR80R for cluster J (Figure 4.1J), snoR66F/119R1, snoR66F/119R2 and 119F1/119R2 for cluster M (Figure 4.1M), snoR114F/snoR115R, snoR114F/snoR85R and snoR115F/snoR85R for cluster N (Figure 4.1N).

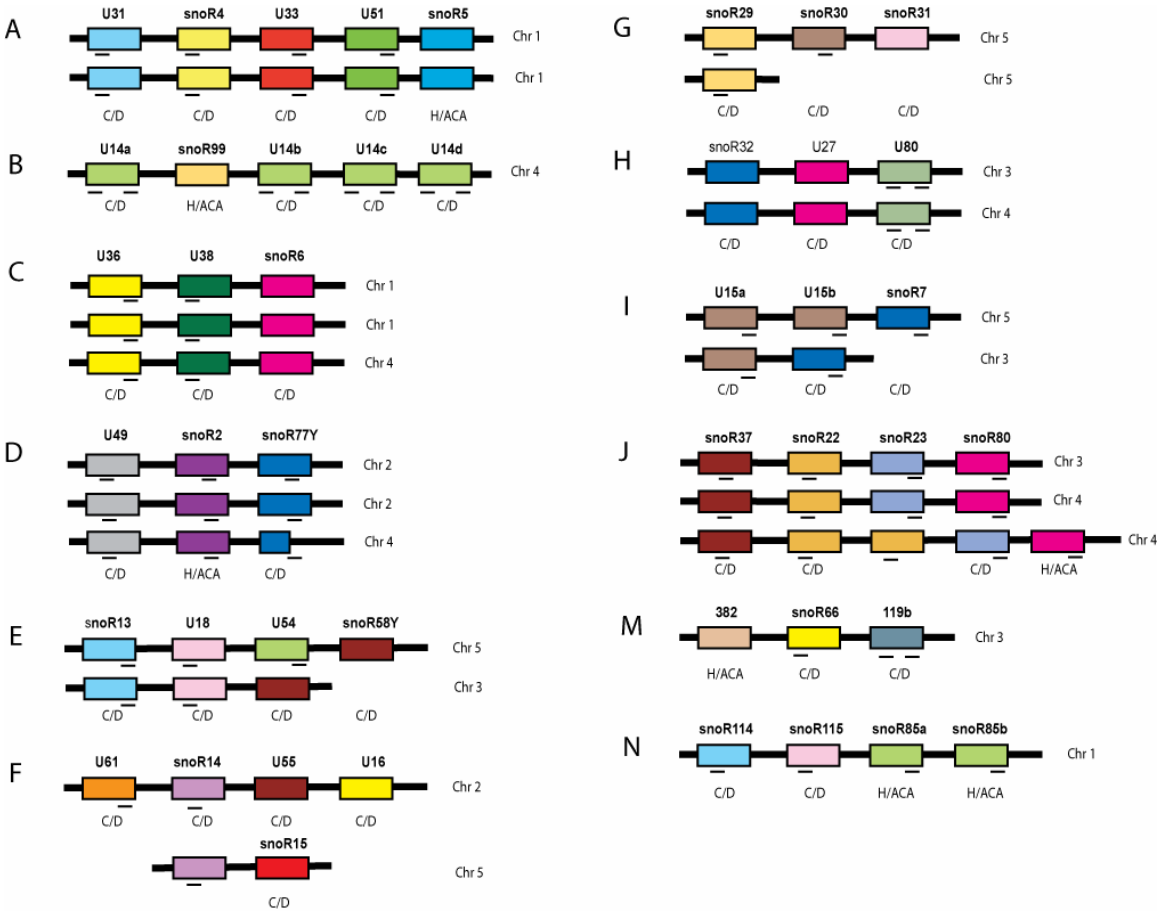


Figure 4.1: Gene organisation in *Arabidopsis thaliana* for the snoRNA genes and gene clusters investigated. Gene clusters are indicated by capital letters (clusters K and L were not investigated and therefore not shown). The approximate location of the universal primer sites are indicated by black lines below the genes. Genes are displayed by boxes of different colours with their names written above and with their chromosome number on the right. The letters C/D and H/ACA positioned below the genes indicate the snoRNA gene type. Note the two different primer sites for the 119b gene (1: forward and reverse primer at the beginning (5' end) and 2: reverse primer at the end (3' end) of the gene).

Usually the forward primers were either radioactively (as in the initial primer-trial) or fluorescence labelled (as in the more detailed examination). PCR amplification was conducted and samples were profiled as described in Chapter 2. In the more detailed investigation, reproducibility of fragment patterns was tested by replicating the entire

procedure for some samples. While few replicates were examined for some primer pairs and no replicates for others, two and three replicates for 34 samples were examined for the U14-3/U14-4 primer combination. Furthermore, most of these replicates were profiled using the different internal size standards ROX500 and ROX1000.

4.2.3 Data scoring

Autoradiographs produced during the initial primer-screen were examined for (i) amplification success, (ii) number of fragments obtained and (iii) fragment-length variation between the samples/species. Presence/absence (1/0) matrices were generated after scoring bands manually.

Raw data obtained from the more detailed investigation of variation were aligned with the internal size standard (ROX500 and ROX1000, Applied Biosystems) and electropherograms were scored using Genemapper 4.0 analysis software (alternatively peakscan; both Applied Biosystems) following the AFLP scoring instructions (Genemapper v3.7 AFLP Analysis). Usually, fragments in the size range of 90 to 800 bp were scored and a presence/absence (1/0) matrix generated. Scoring was performed automatically and checked manually; only peaks with heights above a certain cut off value (usually about 5 to 10 % of the highest peak or the sum of all peaks between 90 and 800 bp, respectively) were positively called. To avoid false calling, peaks that were only one bp apart (double peaks) were usually scored as a single peak (double peak fusing).

4.2.4 Quantifying error rate

A high number of replicates was examined for the U14-3/U14-4 fluorescence labelled primer pair to quantify the error rate for scoring fragment profiles. First, the similarity value (SV) between two replicate profiles of the same sample was calculated by dividing the number of mismatches (nM) between the two profiles by the total number of different peaks (nP) found in these profiles (nM/nP). When three replicated profiles were available each possible profile pair was compared and averaged. Secondly, all observed similarity scores were summed and divided by the samples (nS) ($\sum$SV/nS) to get an average similarity value (ASV). Lastly, the error rate was obtained by subtracting the ASV from 1

(1-ASV). Error rates were calculated for the U14-3/U14-4 dataset with and without double peak fusing.

4.2.5 Analysis of fragment frequencies

Tables containing fragment frequencies were produced from the 1/0 matrices obtained by the fluorescence labelled fragment analysis (Figure 4.2A) by calculating the frequencies of all fragments of different sizes (fds, equivalent to fragments in AFLP datasets) obtained (Figure 4.2B). Columns that did not show frequencies of more than 0.29 within at least one species were deleted (Figure 4.2C). According to their frequencies, fragments were placed into three categories: fragments with frequencies equal to or greater than 0.5 (high frequency fragments, hffs) were shaded grey, whereas fragments with frequencies of 0.3 to 0.49 (moderate frequency fragments, mffs) and fragments with frequencies below 0.3 (low frequency fragments, lffs) were left unshaded (Figure 4.2). These tables which usually consist of a subset of fds were examined for species specific fragments (fragments present in only one species at high frequency) and hffs shared by hybrids and parents. When examining hybrid-parent relationships, only fragments present at high frequency in the hybrid and one parent, but absent or present at very low frequency in the other parent were of interest.

A

Species	Sample	fds				
		1	2	3	4	5
A	1	1	0	1	0	0
	2	0	0	1	1	0
	3	0	0	1	1	0
	4	0	0	1	1	0
	5	0	0	1	1	1
B	1	1	0	0	0	0
	2	1	0	0	0	1
	3	1	0	0	0	0
	4	1	1	1	0	0
	5	0	1	0	0	0
C	1	1	0	0	1	0
	2	1	0	0	1	0
	3	1	0	0	1	0
	4	1	0	0	0	1
	5	1	0	0	1	0

B

Species	fds				
	1	2	3	4	5
A	0.2	0	1	0.8	0.2
B	0.8	0.4	0.2	0	0.2
C	1	0	0	0.8	0.2

C

Species	fds			
	1	2	3	4
A	0.2	0	1	0.8
B	0.8	0.4	0.2	0
C	1	0	0	0.8

Figure 4.2: Construction of fragment frequency tables. From the original 1/0 data matrix (A), the frequencies of all fragments of different size (fds) within a species are calculated and a new data matrix is produced (B). After removing all fds with frequencies less than 0.3 within each species (i.e. fds 5) a fragment frequency table (C) is obtained and examined for putative 'species specific' and possibly shared hybrid-parent fragments. Fragments with frequencies equal to or greater than 0.5 (high frequency fragment, hff) are shaded in grey.

4.2.6 Molecular data analyses

Analyses of molecular data were conducted on each primer pair dataset that showed variation in fragment profiles among any species (initial primer-trial investigation) or among *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis* (i.e. in the more detailed analysis). Some of these variable datasets were also combined and analysed. The combined data sets were constructed after either removing samples that were not present in all fragment datasets (pruned - P datasets) or following the introduction of missing data (MD datasets). The datasets produced from the initial primer-trial were only subjected to individual based analyses through the generation of neighbour joining trees and sometimes Principal Coordinate (PCO) plots (see below).

4.2.6.1 Missing data

Missing data were treated by pairwise deletion (e.g. in PAST) where samples are excluded from any calculation for which they have missing data. Alternatively, missing data were interpolated by sample-by-sample pairwise distances (e.g. in GenAlex) where the average genetic distances for each group level were inserted.

4.2.6.2 Genetic distance analysis – Neighbour Joining (NJ) and Principal Coordinate (PCO) analyses

The Neighbour Joining (NJ) cluster method minimizes the total length of the phylogram by sequentially grouping similar OTUs (operational taxonomic units) (Saitou & Nei, 1987).

The Principal Coordinate Analysis (PCO) finds eigenvalues and eigenvectors of a distance or similarity matrix between all data points and the relationship between the data points can be visualized in a low dimensional space reflecting the original distances as well as possible. PCO is normally performed in three steps. Firstly, a similarity/distance matrix of all data points is produced. Secondly, the matrix is double-centred summing all columns and rows to zero. Thirdly, the transformed matrix is factored and an eigen analysis is performed. The eigenvectors are normalised and the sum of squares of its components equals the corresponding eigenvalues.

The elements of the normalised eigenvectors are the coordinates of the data points representing exactly the distance between them in multidimensional space. The coordinates are adjusted relative to their rectangular and independent principal axis. Thus, the first dimension accounts for the greatest amount of variance and each subsequent dimension explain progressively less of the variance.

NJ and PCO analyses were conducted on matrices of dice similarity index which puts more weight on the joint occurrences of fragments than on shared absence. For combined datasets, bootstrap values (Felsenstein, 1985) for the NJ trees were obtained using 1000 pseudoreplicates. All forms of analysis were conducted using the software PAST 1.99 (Hammer *et al.*, 2001).

4.2.6.3 Genetic distance between NJ trees

To investigate whether datasets generated for different snoRNA genes and gene clusters across the same set of taxa contained similar phylogenetic information, distances between trees for each single dataset used in combined data analyses, and the combined matrix containing all datasets, were calculated using TREEDIST implemented in the PHYLIP package version 3.67 (Felsenstein, 2007). The Branch Score Distance (Kuhner & Felsenstein, 1994) was used in calculations because it takes into account branch lengths. Only datasets consisting of the same samples were used in these analyses and therefore a NJ tree for each single primer pair matrix of the combined and pruned dataset, each containing 43 samples, was produced in PAST 1.99. The NJ trees obtained in Newick notation were copied into a single file which was processed using TREEDIST (with option 2 changed: full distance matrix of distances between all possible trees) and the distance matrix was used for PCO analysis in GenAlEx 6.3 (Peakall & Smouse, 2006).

4.2.6.4 Analyses of molecular variance (AMOVA)

This statistical procedure is used to partition genetic variation at different hierarchical levels (e.g. among individuals within populations, among populations within a region and between different regions). It was initially developed for RFLP haplotypes (Excoffier *et al.*, 1992) but can also be used for many other markers. For binary data, pairwise genetic distances can be estimated using the Euclidean distance metric of Huff et al. (1993). The significance of the variance components can be tested by random permutation.

To quantify levels of genetic differentiation within and among (groups of) species estimates of variance components were assessed by analyses of molecular variance (Excoffier *et al.*, 1992) performed in GenAlEx 6.3 (Peakall & Smouse, 2006). Species were grouped into 'species groups' based on the results of the genetic distance analyses, phylogenetic relationship and ploidy level. Therefore, *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* (closely related diploids) were put in one 'species group', *S. vulgaris* and *S. cambrensis* (tetra/hexaploid and *S. vulgaris* is more distantly related) in another 'species group' and *S. madagascariensis* (distant relative), when available, into a third group. Other species could not be included because of low number

of samples available. However, analyses were performed with one, two (*S. vulgaris*, *S. cambrensis* and *S. madagascariensis* grouped together), and three 'species groups'. Furthermore, for datasets containing *S. madagascariensis* additional analyses without this species were also carried out.

For combined datasets, pairwise Φ_{ST} values (analogous to Fisher's F_{ST} values) were estimated to measure differentiation between species. Furthermore, separate AMOVAs for each species, except *S. madagascariensis*, were conducted. Due to the low numbers of individuals per population, some populations were excluded from analysis, while others were assigned to populations in the same area, and a few were geographically grouped. For example, only one sample of *S. vulgaris* from the population in Egypt was available and was, thus, removed. The only *S. squalidus* sample from the Summerhill population was assigned to the population from Pentre and all *S. cambrensis* individuals from different populations in Wales were treated as one population.

4.2.6.5 STRUCTURE assignment tests

The genetic structure of all variable primer pair matrices was analysed by a model based clustering approach implemented in the computer programme STRUCTURE 2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2007; Hubisz *et al.*, 2009) which can handle dominant markers by introduction of a recessive allele. A single fragment of different size (fds) observation (i.e. one column in the datamatrix) consists of presence (1) or absence (0) of a fragment. Absence of a fragment is the recessive state whereas the presence of fragment represents an ambiguous underlying genotype (in diploids: 11, 10 and 01, respectively). According to its probability, one of these ambiguous genotypes is randomly chosen in each iteration (Falush *et al.*, 2007). This programme is able to calculate the probability P(X|K) for different numbers of natural genetic groups (K) which are distinguished by allele frequencies using a Bayesian algorithm in combination with a Markov Chain Monte Carlo (MCMC) simulation. STRUCTURE analyses were performed for each variable data set and their subsets (e.g. *S. cam* datasets) with K set from K =1 to K = 9 (with 5 replicates for each K), assuming no-admixture model and uncorrelated allele frequencies using a burn-in period of 20000 and 50000 MCMC

repeats. These settings (burn-in and MCMC values) were long enough to stabilize log alpha and Ln likelihood (burn-in) and to obtain consistent end results (MCMC) (Pritchard *et al.*, 2000). Three functions, "Structure.deltaK", "Structure.Table" and "Structure.simil", of the R-script STRUCTURE-SUM-2009.R (Ehrich, 2006; Ehrich *et al.*, 2007) were chosen to decide which K-value and STRUCTURE run would best explain the data. The former function generated 4 plots (Mean L(K), Mean L'(K), Mean L''(K) and Mean DeltaK, respectively) for the determination of the number of groups (K) using the method described in Evanno *et al.* (2005). The number of groups within the plots was indicated by a more or less clear break (plots Mean L(K) and Mean L' (K)) and peak in the slope (plots Mean L''(K) and Mean DeltaK), respectively. However, the most reliable indication of the real K value was shown by the modal value of the Mean DeltaK distribution and its hight might be used as a parameter for the strenght of the signal (Evanno *et al.*, 2005).

Alternatively, the number of groups were chosen using the latter two functions. "Structure.Table" plots the likelihood of each K value (lnP), while "Structure.simil" estimates and plots the similarity among the results of all replicates for each K. The number of groups (K) was chosen when either the lnP in the "Structure.Table" plot showed a maximum or the curve started to even out, the replicates displayed highest similarity ("Structue.Table" and "Structure.simil" plots), and no empty groups were obtained. The run displaying the highest lnP was taken from barplot outputs (see Nordborg *et al.*, 2005) which were further examined to confirm the number of groups.

The ancestry of *S. squalidus* and *S. cambrensis* samples was estimated according to the admixture model by assuming that all hybrid individuals were derived from two populations representing their parents (i.e. *S. squalidus*, *S. aethnensis* and *S. chrysanthemifolius*; *S. cambrensis*, *S. squalidus* and *S. vulgaris*). The clustering procedure determines the proportions of an individual's ancestry derived from these populations (Pritchard *et al.*, 2000). STRUCTURE analyses were performed for combined datasets containing hybrid and parents samples, the latter were predefined (USEPOPINFO = 1), with K set to 2 (with 5 replicates) using a burn-in period of 20000 and 50000 MCMC repeats. The run with the highest lnP was taken from barplot outputs.

4.3 Results

4.3.1 Radioactively labeled fragment analysis (initial primer-trial investigation)

Amplification was successful in the majority of samples using the following eleven primer combinations U14-1/U14-2, U14-3/U14-4, U33F/U51R, U31F/U33R, U31F/U51R, snoR13F/U18R, U18F/U54R, U52F/snoR22R, snoR22F/snoR23R, U61F/snoR14R, snoR29F/snoR30R and snoR30F/U34R. Products from two primer pairs, U49F/snoR2R and snoR2F/snoR77YR, were amplified in only a few samples, and five primer-pairs, U31F/snoR4R, snoR4F/U33R, U36F/U38R, U80F/U80R and U15F/snoR7R, failed to amplify a product in any sample. The data produced for each primer pair were subjected to NJ analysis and an examination of fragments shared between hybrid species and parents. However, only the results for the snoR29F/snoR30R primer pair are presented in detail here. As this is an initial investigation, figures illustrating results obtained for other primer combinations are not presented in this thesis but are available in electronic format on the accompanied CD (supplemental material).

4.3.1.1 snoR29F/snoR30R primer pair

Twenty different fragment profiles were produced across 26 samples examined in the initial primer-trial using the snoR29F/snoR30 primer combination. Fragment sizes ranged from 205 to 260 bp in length and were therefore shorter than the size recorded in *A. thaliana* (284 bp) (Figure 4.3). The number of fragments per sample varied between one (e.g. *S. madagascariensis* (22)) and six (e.g. *S. massaicus* (6)).

Fragment profiles varied between species and different fragment patterns were found between samples within species apart from within *S. flavus* (Figure 4.3) A high level of fragment pattern variation was evident within *S. glaucus* where all samples differed in number and sizes of their profiles and only one fragment (216 bp) was shared between *S. glaucus* samples (5) and (17) (Figure 4.3).

A number of fragments were shared between hybrids and their parent species (Figure 4.3). While *S. squalidus* (sample 3) shared only one band (227 bp) with one

parent, *S. aethnensis* (sample 1), three fragments (216, 220 and 227 bp) of *S. squalidus* (sample 10) were shared with its parents (*S. aethnensis* (1): 216 and 227bp and *S. chrysanthemifolius* (2): 220bp). However, three of the *S. squalidus* (10) fragments (216, 220 and 250) were also present in *S. vulgaris* (7) and (8), the latter sharing three bands (216, 220 and 243 bp) with its hybrid *S. teneriffae* (11). *S. cambrensis* (9) shared two fragments (216 and 250 bp) with each of its parents, *S. squalidus* (10) and *S. vulgaris*. The bands seen in *S. flavus* (at 216, 226 and 260 bp) were also present in its hybrid *S. mohavensis* (19, 20, 21) (Figure 4.3). Due to low number of samples examined for each species the detection of hybrids based on shared fragments between hybrid species and one of its parents is of limited value. However, the results obtained here might be useful for choosing regions for a more detailed investigation.

Samples	Species	snoR29/snoR30 fragment sizes (bp)																	
		205	208	212	213	214	216	220	221	222	224	226	227	240	243	248	249	250	260
1	*S. aethnensis*						■						■						
2	*S. chrysanthemifolius*		■					■											
3	*S. squalidus*	■			■														
10	*S. squalidus*						■	■					■					■	
7	*S. vulgaris*						■	■										■	
8	*S. vulgaris* var. *hib.*						■	■							■			■	
9	*S. cambrensis*						■		■	■								■	
11	*S. teneriffae*						■	■							■				
12	*S. teneriffae*				■				■				■						
5	*S. glaucus*	■					■	■		■	■		■						
16	*S. glaucus*					■									■				
17	*S. glaucus*				■		■												
4	*S. massaicus*		■						■					■		■		■	
6	*S. massaicus*		■	■					■					■		■		■	
13	*S. flavus*						■												■
14	*S. flavus*						■												■
15	*S. flavus*						■												■
19	*S. mohawensis* ssp. *bre.*						■			■		■							■
20	*S. mohawensis*						■			■		■							■
21	*S. mohawensis*						■					■							■
22	*S. madagascariensis*																	■	
23	*S. madagascariensis*																	■	
24	*S. madagascariensis*																	■	
25	*S. madagascariensis*																■	■	
26	*S. madagascariensis*																■		■
27	*S. madagascariensis*																■	■	

Figure 4.3: Fragment profiles of *Senecio* ssp. generated using primers snoR29F/snoR30R. Twenty different fragment profiles were generated among 26 samples surveyed across 13 species/subspecies. The matrix shows fragments obtained for each sample. *hib. = hibernicus*; *brev. = breviflorus*.

In the NJ tree, some samples clustered according to species. For example, all *S. madagascariensis* samples, except *S. madagascariensis* (26), were placed in the same cluster. *S. madagascariensis* (26) shared one of its two fragments with *S. mohavensis* and *S. flavus* and was placed within a cluster containing these species. However, for some species, different samples were placed in different clusters within the tree. For instance, samples of *S. squalidus*, *S. teneriffae* and *S. glaucus* were present in two different clades at the base of the tree. The species within these two clades are intermixed and no clear hybrid-parent relationship is seen. Furthermore, *S. cambrensis* is not found within these clades, but is placed at the base of the clade containing most of the *S. madagascariensis* samples. However, the hybrid species *S. mohavensis* is placed with *S. flavus*, which is one of its parents.

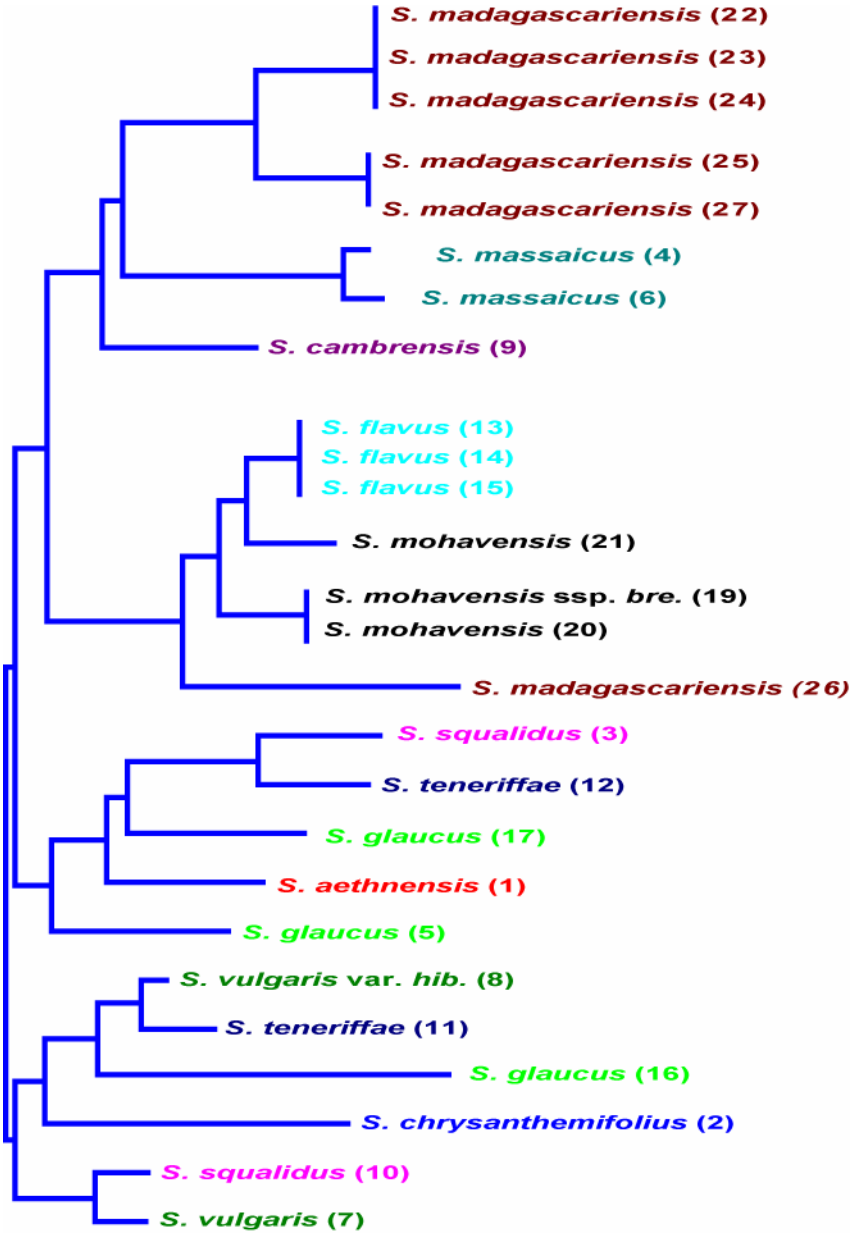


Figure 4.4: NJ tree of *Senecio* sp. based on fragment variation generated by snoR29F/snoR30R primers. (*hib.* = *hibernicus*; *brev.* = *breviflorus*).

4.3.1.2 Summary of all clusters

Most primer pairs that successfully amplified products in the initial primer-screen revealed variation between and within species and species groups (i.e. the *S. squalidus* group plus *S. vulgaris* and *S. teneriffae*) (Table 4.1). However, the amount of variation between and within species depended on the cluster/primer pair examined. For example, *S. flavus, S. mohavensis* and *S. madagascariensis* often possessed distinct fragment profiles and formed clusters according to their species within the NJ trees (e.g. U33/U51 U14-1/U14-2 and U14-3/U14-4, snoR13/U18). While the variation within some species appeared to be low to moderate, variation in other species seemed unexpectedly high, with samples from the same species being placed in very different positions in a NJ tree. For example, *S. flavus* showed no within species variation apart from when the primer combinations U31/U51, U33/U51 and U14-3/U14-4 were used, which generated low variation in the species. Similarly, *S. madagascariensis* showed only moderate variation with the primer combinations U33/U51, U14-1/U14-2, U14-3/U14-4 and snoR13/U18. In contrast, all primer combinations generated high levels of variation within *S. glaucus*. Within the *S. squalidus* group of species, variation between species seemed as high as within species, and variation was generated by all primer combinations except U33/U51.

Fragments shared between at least one parent and their hybrid were generated by all primers except snoR2/snoR77Y (Table 4.1). For example, all fragments generated in *S. flavus* were also present in *S. mohavensis* for all primer combinations except U33/U51 and U14-3/U14-3. For these last two primer sets *S. flavus* possessed some fragments that were not found in *S. mohavensis*. As expected, tetraploid *S. mohavensis* contained additional fragments to those found in diploid *S. flavus*, however these additional bands could not be assigned clearly to its other diploid parent, *S. glaucus*, because this species was highly variable in fragment profile and only a few samples of it were examined. Within the *S. squalidus* group of species, which consisted predominantly of hybrids and their parents, fragments were shared between species. However, due to a lack of fragments that were exclusively shared between parents and hybrids, it was difficult to identify bands that could be used satisfactorily for hybrid detection. Within this group *S. squalidus* was both a parent of a hybrid (i.e. *S. cambrensis*), and also a hybrid itself (i.e. of *S. aethnensis* and *S. chrysanthemifolius*), and it appeared that one sample of *S.*

squalidus (10) shared more bands with its hybrids (*S. cambrensis* and *S. teneriffae*), whereas the other sample of *S. squalidus* (3) was closer in fragment type to its parents (*S. aethnensis* and *S. chrysanthemifolius*). Most primer combinations generated fragments in hybrids that were not present in their parents and vice versa. Thus, hybrid detection using snoRNA markers is likely to be more successful if based on shared fragments than on genetic distances computed from fragment profiles.

In the diploid *Senecio* species tested, it was estimated that the number of copies of each snoRNA cluster (Figure 4.1) were as follows (Table 4.1): at least three copies of cluster A (based on five bands generated by U31/U51 and U33/U51), two copies of cluster D (based on two U49/snoR2 bands of 410 and 550 bp), three copies of cluster E (five bands generated by snoR13/U18), two copies of cluster F (three bands generated), two copies of cluster G (three bands generated) and three copies of cluster J (five bands generated by snoR37/snoR22 in *S. squalidus*). Cluster B consisted of homologous genes only and the number of gene cluster copies could not be estimated. However, the primer combination U14-3/U14-4 generated six different fragments, the same number of fragments was obtained by reverse ePCR for *A. thaliana* suggesting a gene copy number similar to this species.

Table 4.1: Summary of radioactively labeled fragment analysis results. 14 primer combinations showed PCR amplification. Amplification success is indicated by the number of samples amplified/number of samples tested. + = present; - = absent. SR = snoR.

Cluster	Primer pair	Amplification success	Max. number of bands in diploids	Variation		shared parents/hybrid fragments
				between species	within species/ species group*	
A	U31/U51	14/15	5	+	+/+	+
	U31/U33	27/28	3	+	-/+	+
	U33/U51	28/28	5	+	+/-	+
B	U14-1/U14-2	24/28	2	+	+/+	+
	U14-3/U14-4	26/28	6	+	+/+	+
D	U49/SR2	15/28	2	+	+/-	+
	SR2/SR77Y	4/28	2	-	-/-	-
E	SR13/U18	25/28	5	+	+/+	+
	U18/U54	24/28	2	+	+/+	+
F	U61/SR14	14/15	3	+	-/+	+
G	SR29/SR30	27/28	3	+	+/+	+
J	SR37/SR22	15/15	9	+	+/+	+
	SR22/SR23	25/28	4	+	+/+	+

* *S. squalidus* group including *S. vulgaris, S. vulgaris* var. *hibernicus* and *S. teneriffae*

4.3.1.3 Combined data

Except for snoRNA cluster D for which only a few samples were successfully amplified, all datasets (i.e. cluster A, B, E, F, G and J datasets) and at least one dataset from each cluster (i.e. U33/U51, U14-3/U14-4, snoR13/U18, U61/snoR14, snoR29/snoR30 and snoR22/snoR23), chosen by the highest number of samples were combined and analysed by introducing missing data (MD datasets). In the NJ tree generated after combining all datasets (Figure 4.5A), samples of *S. mohavensis, S. flavus, S. madagascariensis* and *S. massaicus* clustered together according to species, with each group except the *S. madagascariensis* one having high bootstrap support. Two of the three *S. glaucus* samples examined grouped together and showed a close relationship with *S. mohavensis* and *S. flavus* in the tree; however, the third sample of the species was more distantly related and was positioned outside the cluster containing *S. mohavensis, S. flavus, S. madagascariensis* and *S. massaicus.* All species of the *S. squalidus* group (including *S. vulgaris, S. vulgaris* var. *hibernicus* and *S. teneriffae*) were positioned at the base of the tree along with *S. aethnensis* and *S. chrysanthemifolius,* with little evidence of distinct

clusters forming between samples according to taxon. With minor exceptions the groupings of samples in the NJ tree produced after combining at least one dataset for each snoRNA cluster (Figure 4.5B) were similar to those in the tree generated after combining all datasets (Figure 4.5A). Principal Coordinate Analyses (PCO) of both combined datasets detected five separated groups in plots of individual scores against the first two principal coordinates (Figure 4.5C and D), which explained more than 40% of the total variation. One group consisted of *S. madagascariensis*, another of *S. flavus*/*S. mohavensis*, a third of *S. massaicus*, a forth of *S. glaucus* (16, 17) and the fifth group members of the *S. squalidus* group (including *S. vulgaris*, *S. vulgaris* var. *hibernicus* and *S. teneriffae*)/*S. glaucus* (5) in both datasets.

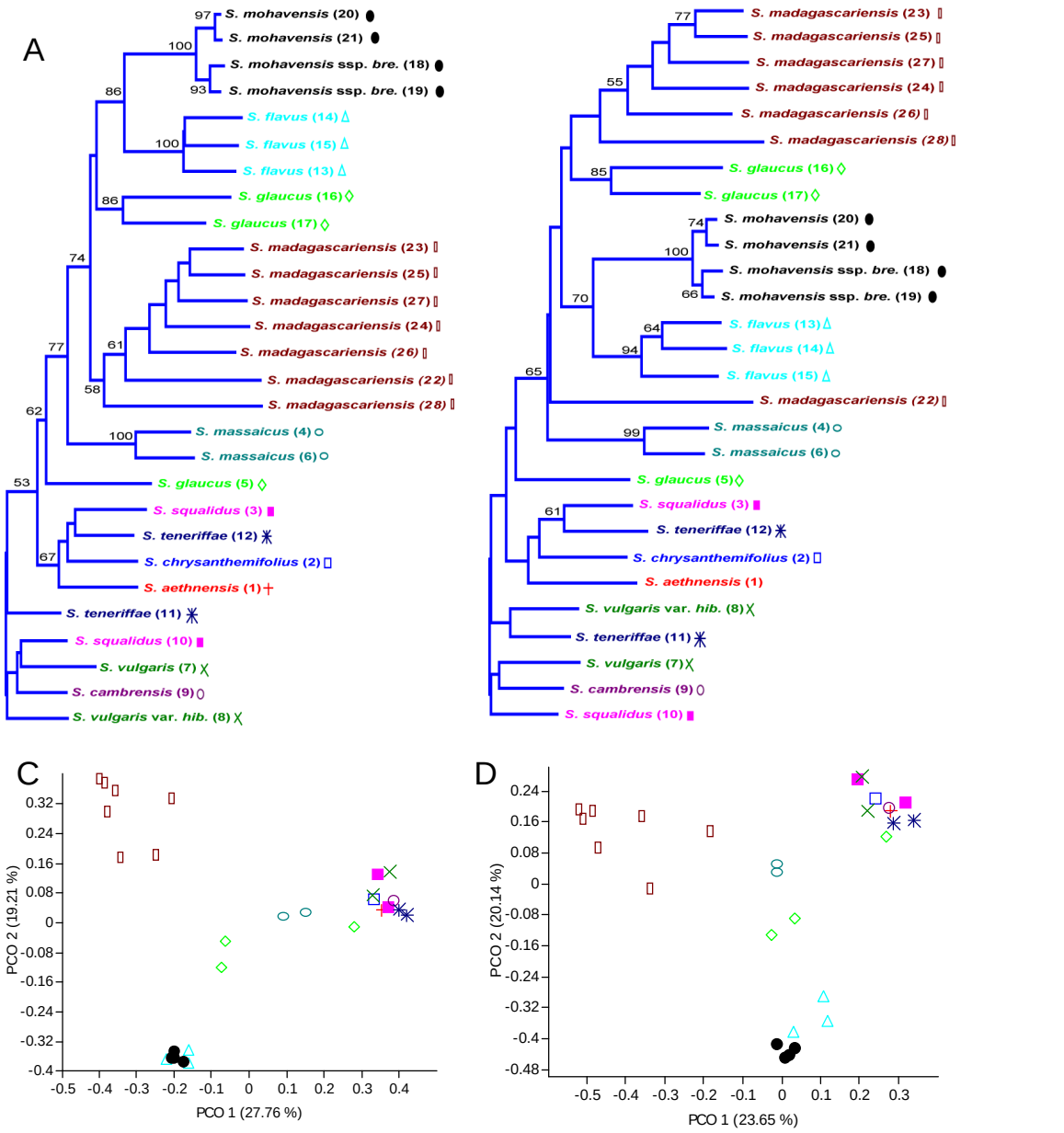


Figure 4.5: NJ trees and Principal Coordinate Analyses (PCO) of combined datasets. A and B: NJ tree and PCO plot for a combined dataset containing all but snoRNA cluster D matrices. C and D: NJ tree and PCO plot for a combined dataset containing only one matrix of each cluster apart from cluster D. For snoRNA gene clusters for which there was more than one dataset available, the one containing most of the 28 samples and showing the most structured NJ tree (see Appendix) was selected for combination. Symbols next to the species in the NJ trees are used in the PCO plots. Bootstrap values (>50 %) based on 1000 replicates are shown above or below branches in the NJ trees. *hib.* = *hibernicus*; *brev.* = *breviflorus*.

4.3.2 Fluorescence labelled fragment analysis (more detailed investigation)

4.3.2.1 Amplification success, putative non-specific fragments and error rate

PCR amplification was successful for 28 primer combinations examined, but five (shaded grey; Table 4.2) were not used for fragment analysis. Only two primer-pairs, U31/SnoR4 and snoR4/U33, both of which included a primer within the putative snoRNA4 gene, did not amplify (Table 4.2, shaded turquoise). Some samples, in particular *S. squalidus* individuals, could not be amplified for every primer pair, despite having high quality DNA and, thus, no fragment profiles were available for these samples.

Due to the low number or lack of replicates in most datasets, the error rate was calculated only for the U14-3/U14-4 primer combination. Furthermore, most samples used for replicates were genotyped with two different internal size standards (ROX500 and ROX1000). Although the ROX1000 had to be adjusted for each single sample, the fragments obtained had the same size for each size standard as expected and, therefore, combining these datasets was possible. The calculated error rate based on 34 samples (38 fragment profiles) comprising all 5 species of the *S. squalidus* group resulted in a value of 0.1 (double peaks scored as 2 peaks) and 0.03 (double peak fusing), respectively. Although there were only a few replicates available for other datasets, most of them showed the same profile topology, but sometimes differed in peak size.

4.3.2.2 Fragment profiles

In total, 1134 fragment profiles (fps) were obtained using 23 different primer combinations from 13 different snoRNA gene clusters (Table 4.2). The majority of profiles belonged to *S. aethnensis* (181 fps), *S. chrysanthemifolius* (197 fps), *S. squalidus* (301 fps), *S. vulgaris* (163 fps) and *S. cambrensis* (169 fps). Additionally, fragment profiles were available for *S. madagascariensis* (65 fps), *S. flavus*, *S. massaicus*, *S. engleranus*, *S. teneriffae* and *S. glaucus* (total of 58 fps, Table 4.2) for some primer

combinations. Some gene clusters were examined using different primer combinations. As all of these genes should have the same evolutionary history, some were examined for only a subset of individual specimens. For example, for gene cluster J, only the snoR22/snoR23 primer pair was used for typing the full sample set. The snoR37/snoR22 and the snoR37/snoR23 primer combinations were investigated using a subset of 58 and 38 samples, respectively. However, with the exception of snoR13/U54 and U33/U51, at least four – though usually more - samples of each of the five species comprising the *S. squalidus* group (*S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis*) were typed.

Table 4.2: Number of samples per species producing fragment profiles across primer combinations. Twenty three primer combinations that amplified 12 different snoRNA gene clusters were used for fluorescence genotyping. Another five primer combinations (shaded grey) were successful in PCR amplification, whereas two (shaded green) were not. The other species ('others') referred to in this table included *S. engleranus*, *S. flavus*, *S. teneriffae*, *S. massaicus* and *S. glaucus*. The number of samples per species available for analysis is placed in brackets. SR = snoR.

Cluster	Primer-combination	*S. aethnensis* (11)	*S. chrysanthemifolius* (12)	*S. saqualidus* (29)	*S. vulgaris* (13)	*S. cambrensis* (12)	*S. madagascariensis* (9)	others (11)	Sum (97)
A	U31/U51	11	11	26	11	11	8	7	85
	U33/U51	10	7	8	2	2	-	1	30
	U31/U33								
	U31/snoR4								
	snoR4/U33								
B	U14-1/U14-2								
	U14-3/U14-4	10	11	27	10	10	9	5	82
C	U36/U38	10	11	15	11	11	-	1	59
D	U49/snoR2d	10	10	10	10	11	-	1	52
	snoR2d/snoR77	6	11	10	5	7	-	-	39
	U49/snoR77	9	11	6	6	6	-	-	38
E	snoR13/U18	10	11	27	11	11	9	9	88
	U18/U54	10	11	18	7	10	9	4	69
	snoR13/U54	5	8	3	5	3	-	-	24
F	U61/snoR14	10	11	27	11	11	9	9	88
G	snoR29/snoR30	10	10	27	11	10	7	9	84
H	U80-1/U80-2	10	11	15	9	10	-	-	55
I	U15/snoR7	10	10	5	4	4	-	-	33
J	snoR37/snoR22	10	11	14	11	11	-	-	57
	snoR22/snoR23	11	12	29	13	12	9	11	97
	snoR37/snoR23	9	11	6	6	6	-	-	38
	snoR37/snoR80								
M	snoR66/119R1	5	4	8	5	6	1	1	30
	snoR66/119R2	5	5	8	7	9	4	-	38
	119R1/119R2								
N	snoR115/snoR85	5	5	6	4	4	-	-	24
	snoR114/snoR85	5	5	6	4	4	-	-	24
	snoR114/snoR115								
Sum		181	197	301	163	169	65	58	1134

The number of fragments produced by most primer pairs differed considerably between diploid samples. While most primer combinations produced profiles containing one to 12 fragments, profiles containing 16 and as many as 21 fragments were obtained for the primer pairs U49/snoR2 and U49/snoR77, respectively (Table 4.3). As well as variation in number of fragments amplified among primer pairs, there was considerable variation in size of fragments amplified. The snoR114/snoR85 primer pair, for example, produced fragments ranging from 97 to 770 bp in size, whereas fragments from 100 to 293 bp were amplified by the U61/snoR14 primer pair (Table 4.3).

Table 4.3: Variation in fragment number and size generated by 21 primer pairs. SR = snoR.

Cluster	Primer pair	Numbers of fragments in diploids	Size of fragments (nt)	Cluster	Primer pair	Numbers of fragments in diploids	Size of fragments (nt)
A	U31/U51	1-7	106-628	G	SR29/SR30	2-8	103-390
	U33/U51	2-3	147-159	H	U80-1/U80-2	1-3	128-321
B	U14-3/U14-4	1-4	123-694	I	U15/snoR7	2-3	92-404
C	U36/U38	2	94-157	J	SR22/SR23	2-5	189-544
D	U49/SR2	2-16	116-635		SR37/SR22	1-9	105-438
	SR2/SR77	1-2	93-157		SR37/SR23	3-7	92-592
	U49/SR77	4-21	100-637	M	SR66/119R1	2-6	100-481
E	SR13/U18	1-6	92-667		SR66/119R2	1-4	97-360
	SR13/U54	1-5	97-742	N	SR114/SR85	2-10	97-770
	U18/U54	2-3	98-320		SR115/SR85	6-12	97-770
F	U61/SR14	2-7	100-293				

The data produced for each primer pair were subjected to an analysis of fragment frequencies, NJ, PCO, AMOVA and STRUCTURE analyses described in Materials and Methods. Only the results of these analyses for the snoR29/snoR30 primer pair are presented in detail here. This primer pair yielded the most complete dataset across samples. Figures and Tables of results for the other primer pairs used are presented in the appendix or supplemental material, however a summary of the major findings using all primer pairs is provided in the text and highlighted in different tables throughout this chapter.

4.3.2.3 Analysis of fragment frequencies

Fragment profiles were produced for 84 samples using the snoR29F/snoR30R primer pair. These contained between two to 11 fragments per profile with fragments ranging in size from 103 to 390 bp. Fragments of moderate and high frequency (mffs and hffs, respectively) fell within a size range of 110 to 250 bp (Table 4.4). Several fragments (129 bp, 212 bp and 216 bp) were present in almost all species but varied in frequency across taxa. For example, the 212 bp fragment was present in all species, but ranged in frequency from being fixed (frequency equalled 1.00) in *S. cambrensis*, *S. aethnensis* and *S. flavus* to occurring at a frequency of 0.29 in *S. madagascariensis*. Some fragments were found within a certain group of species, for example two fragments (240 bp and 246 bp) were present exclusively within *S. vulgaris*, *S. cambrensis* and *S. madagascariensis.* Other fragments were present in a particular species, but were absent or rare in other species (e.g., 160 bp; 179 bp; 198 bp; 203 bp; 209 bp; 220 bp; 230 bp and 250 bp; Table 4.4). Thus, the 203 bp and 220 bp fragments were present exclusively in *S. squalidus* and *S. vulgaris*, respectively, where they occurred at high to moderate frequencies, respectively. Fragments of 179 and 198 bp were shared by two species only, the allopolyploid *S. cambrensis* (hybrid) and one of its parents. The former fragment was present in all samples of *S. cambrensis* and *S. vulgaris* examined, whereas the 197 bp fragment was present in *S. cambrensis* (frequency of 0.50) and its other parent *S. squaldius* (frequency of 0.82; Table 4.4).

Table 4.4: Frequencies of fragments amplified by the snoR29/snoR30 primer pair within each species. Only fragments with a frequency of at least 0.33 within species (moderate frequency fragments (mffs)) are shown. Within species frequencies above 0.5 (high frequency fragments (hffs)) are shaded in grey.

		SR29/SR30														
Species	**N**	**110**	**129**	**160**	**179**	**198**	**201**	**203**	**209**	**212**	**216**	**220**	**230**	**240**	**246**	**250**
S. aethnensis	11	0.36	0.27	0.09			0.27	0.18		1.00	1.00		0.82			
S. chrysanthemifolius	10		0.10				0.10	0.90		0.50	1.00		0.40			
S. squalidus	28	0.29	0.36			0.82	0.36	0.25	0.64	0.71	0.75		0.21			
S. vulgaris	11	0.18	0.36	0.09	1.00					0.27	1.00	0.36		0.64	0.55	
S. cambrensis	10	0.50	0.80		1.00	0.50				1.00	1.00			0.90	0.50	
S. madagascariensis	7	0.29	0.14	0.57			0.43	0.71		0.29	0.14		0.14	0.71	1.00	
S. flavus	3									1.00						1.00

Other high frequency fragments shared between a hybrid taxon and either one or the other parent taxon were resolved by most primer pairs except U33/U51, U14-3/U14-4, U36/U38, snoR2/snoR77, U80-1/U80-2, U15/snoR7, U49/snoR77, snoR13/U18 and snoR13/U54. Most of these were identified in comparisons between *S. cambrensis* and *S. vulgaris* (C-V, 20) and between *S. cambrensis* and *S. squalidus* (C-S, 12) and only a few were identified as shared between *S. squalidus* and its parents *S. chrysanthemifolius* (S-Ch, 6) and *S. aethnensis* (S-A, 2) (Table 4.5 and Table 4.6).

Table 4.5: Fragments shared between hybrid species and parent taxa generated by 12 primer combinations.

Species	*S. aethnensis*	*S. chrysanthemifolius*	*S. squalidus*	*S. vulgaris*
S. squalidus	2	6	-	-
S. cambrensis	-	-	12	20

A summary of fragments identified as shared between hybrid species and their parent taxa is presented in Table 4.6 (for more detailed results see appendix 4, Table A.3 to Table A.20). The U61/snoR14 primer pair amplified one fragment shared between *S. cambrensis* and *S. vulgaris*, whereas the U31/U51 primer combination amplified two such fragments (Table 4.6).

The U36/U38, snoR2/snoR77 and U15/snoR7 primer pairs produced fragment profiles that did not vary across taxa, while some other primer pairs, though generating low levels of variation, nonetheless yielded some species specific fragments (Table 4.6). For example, the U14-3/U14-4 fragment profiles of *S. madagascariensis* and *S. engleranus* included two to three fragments that were not present in other species' profiles. Thus, this primer pair makes it possible to distinguish between *S. madagascariensis*, *S. aethnensis* and *S. flavus*, but does not allow distinction between *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris*, *S. cambrensis*, *S. teneriffae*, *S. massaicus* and *S. glaucus*. Similarly, the U33/U51 primer pair produces the same fragment profile in all species of the *S. squalidus* group, but distinguishes these species from *S. engleranus*. In addition, the U80-1/U80-2 primer pair produces a fragment profile containing one fragment specific to *S. aethnensis*. Other primer combinations, U49/snoR77, snoR13/U18 and snoR13/U54, generated more variable fragment profiles containing fragments specific to *S. cambrensis*, *S. madagascariensis* and *S. chrysanthemifolius*, respectively. Additional species' specific fragments were generated by U31/U51, U61/snoR14 and snoR22/snoR23 for *S. madagascariensis*, snoR37/snoR23 for *S. cambrensis*), snoR114/snoR85 for *S. vulgaris* and *S. chrysanthemifolius* and snoR115/snoR85 for *S. vulgaris*.

Table 4.6: Species specific fragments, and fragments shared between hybrid taxa and parent taxa across 21 primer pairs tested. SR = snoR; V = *S. vulgaris*, C = *S. cambrensis*, Ch = *S. chrysanthemifolius*, Sq = *S. squalidus*, A = *S. aethnensis*; mff = moderate frequency fragment, hff = high frequency fragment. Note that the first letter refers to the hybrid taxon and the second one to the parent taxon that shares a particular fragment.

Cluster	Primer pair	Putative species specific fragments*	Putative hybrid-parent fragments**	Cluster	Primer pair	Putative species specific fragments*	Putative hybrid-parent fragments**
A	U31/U51	M	C-V	G	SR29/SR30	S, O	C-V, C-S
	U33/U51	O	-	H	U80-1/U80-2	A	-
B	U14-3/U14-4	M, O	-	I	U15/snoR7	-	-
C	U36/U38	-	-	J	SR22/SR23	M	C-V, C-S, S-Ch^
D	U49/SR2	-	C-V, C-S		SR37/SR22	-	C-V, C-S
	SR2/SR77	-	-		SR37/SR23	C	C-S
	U49/SR77	C	-	M	SR66/119R1	-	C-V, C-S, S-Ch^
E	SR13/U18	M	-		SR66/119R2		C-S
	SR13/U54	Ch	-	N	SR114/SR85	V, Ch	C-V, S-Ch^, S-A^
	U18/U54	-	C-V, C-S^		SR115/SR85	V	C-V, C-S, S-Ch^
F	U61/SR14	M	C-V^				

* fragments with high frequencies exclusively present in only one species
** fragments which show high frequencies in the hybrid and one parent and are absent (0 to 5 % frequency) or present in low (< 30 %) frequencies (^) in the other parent.

The primer pairs U33/U51, U14-3/U14-4, U36/U38, snoR2/snoR77, U80-1/U80-2 and U15/snoR7 generated fragment profiles that exhibited very low variation across all samples and were not subjected to further analysis of fragment length variation.

4.3.2.4 Neighbour joining trees and Principal Coordinate analyses

The Neighbour Joining (NJ) tree generated from genetic distances between 86 samples analysed using the snoR29/snoR30 primers (Figure 4.6) showed that all samples of *S. vulgaris, S. cambrensis* and *S. engleranus* comprised a single cluster, while six of seven *S. madagascariensis* samples were grouped together with the single *S. massaicus* sample in another cluster. Samples of *S. chrysanthemifolius* and *S. squalidus* tended to group together, although two samples of *S. squalidus* grouped with *S. flavus* while one sample grouped with *S. teneriffae* (o12). Other remaining samples of *S. aethnensis, S. chrysanthemifolius, S. squalidus* as well as *S. glaucus* (o5) were distributed at the bottom of the NJ tree (Figure 4.6).

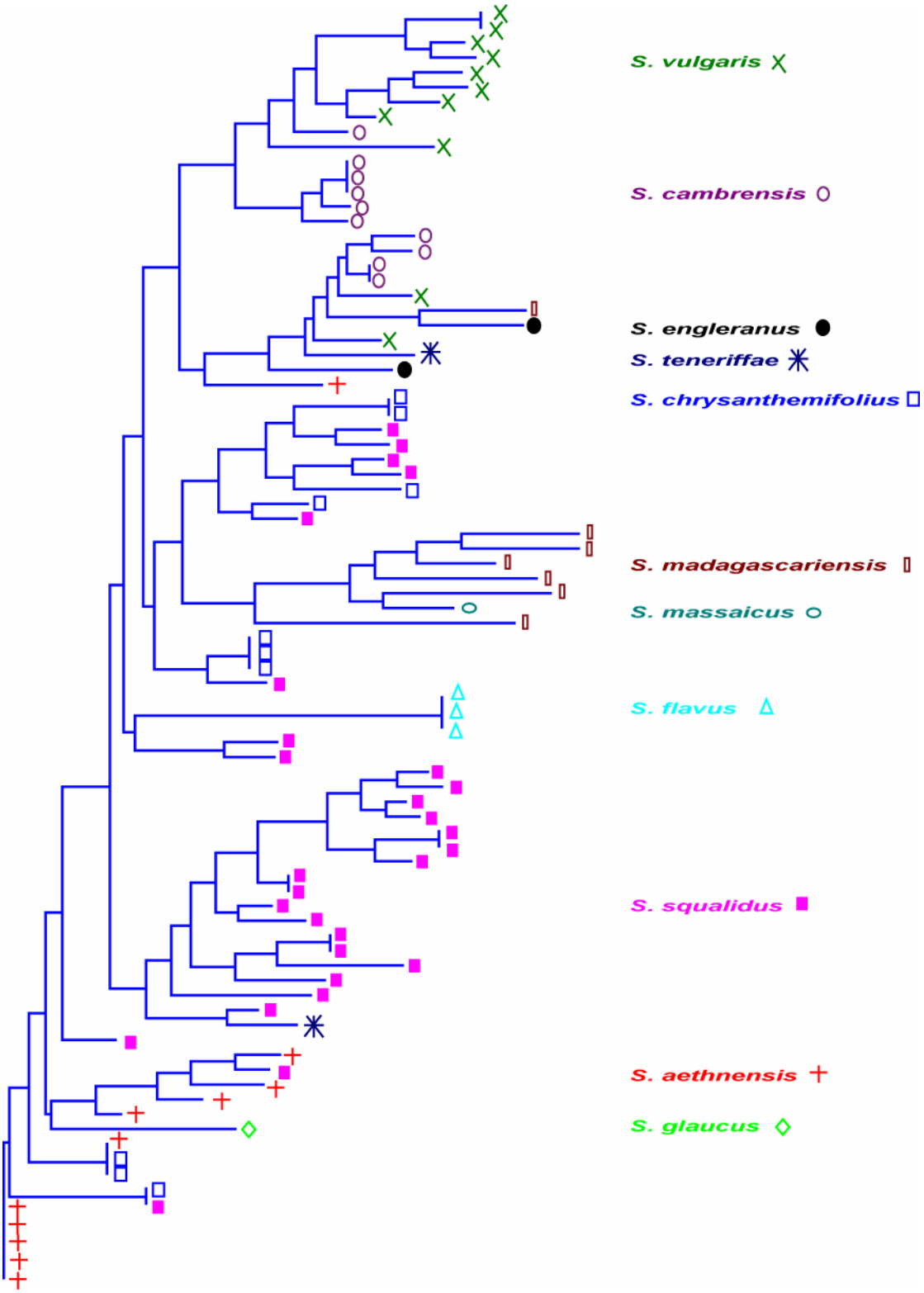


Figure 4.6: NJ tree of *Senecio* sp. based on snoR29/snoR30 fragment profiles. NJ analysis of 84 samples across 11 species is based on fragment variation (31 fds) and dice genetic similarities.

The PCO plot (Figure 4.7) displayed two groups containing various species. One group consisted of *S. squalidus*, which appeared to be highly variable and overlapped with both of its parent species, *S. aethnensis* and *S. chrysanthemifolius*, *S. flavus* and samples of *S. glaucus*, *S. engleranus* and *S. teneriffae*. The other group contained *S. vulgaris*, which overlapped with its hybrid species, *S. cambrensis*, and *S. madagascariensis*, and individuals of *S. massaicus*, *S. engleranus* and *S. teneriffae*. It should be noted that some samples of *S. cambrensis* were placed in close proximity to its other parent *S. squalidus* within the plot (Figure 4.7).

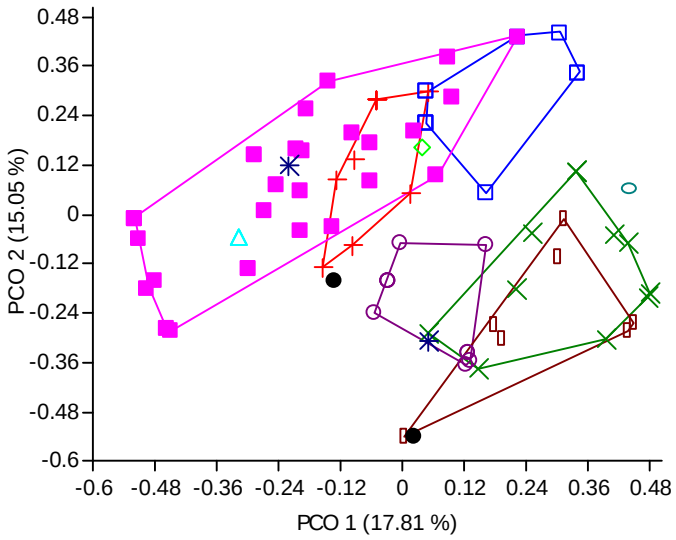


Figure 4.7: PCO plot of *Senecio* sp. snoR29/snoR30 fragment profiles. *S. aethnensis* = +; *S. chrysanthemifolius* = □; *S. squalidus* = ■; *S. vulgaris* = x; *S. cambrensis* = ○; *S. madagascariensis* = □; *S. flavus* = Δ; *S. engleranus* = ●; *S. teneriffae* = *; *S. glaucus* = ◊; *S. massaicus* = ○.

In PCO plots based on distance matrices generated by other primer pairs, *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* samples were never clearly separated into distinct clusters according to taxon. This is not surprising given their very close relationship. However, the taxa were separated, albeit with some overlap, by some primer pairs, for example, U49/snoR2 (Figure 4.8).

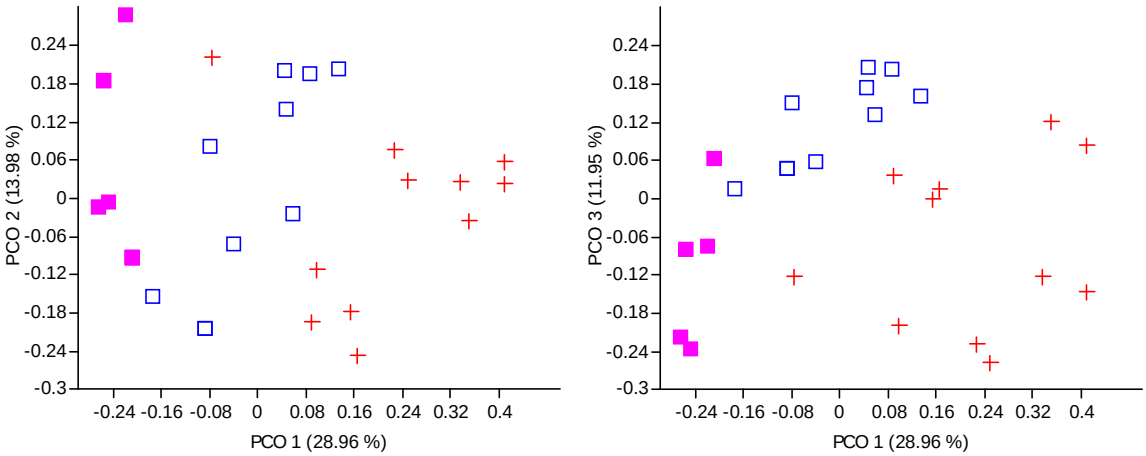


Figure 4.8: PCO plots of *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* fragment profiles. PCO analysis is based on fragment variation generated by the U49/snoR2 primer pair. The first three axes of the PCO explained 54.89 % of the variation within the dataset. *S. aethnensis* (+); *S. chrysanthemifolius* (□); *S. squalidus* (■).

Senecio vulgaris and *S. cambrensis* were grouped together according to fragment profiles generated by each of 12 primer pairs, while *S. madagascariensis* samples, when included in analyses, were usually separated from these taxa (Table 4.7 and appendix 4, Figure A.4 to Figure A.15).

Table 4.7: **Groupings of species identified by different primer pairs**. SR = snoR; V = *S. vulgaris*, C = *S. cambrensis*, V/C = *S. vulgaris* and *S. cambrensis* clustered, CR = closely related species, M = *S. madagascariensis*, Ch = *S. chrysanthemifolius*, Sq = *S. squalidus.*

Cluster	Primer pair	Groupings	Cluster	Primer pair	Groupings
A	U31/U51	V/C, M	J	SR22/SR23	V/C, M
D	U49/SR2	CR		SR37/SR22	V/C, CR
	U49/SR77	V/C, M		SR37/SR23	V/C, CR
E	SR13/U18	M	M	SR66/119R1	V/C, CR
	SR13/U54	C, Ch		SR66/119R2	V, M, CR
	U18/U54	V, DR	N	SR114/SR85	V/C
F	U61/SR14	V/C, M		SR115/SR85	V/C, CR
G	SR29/SR30	V/C, M, Sq			

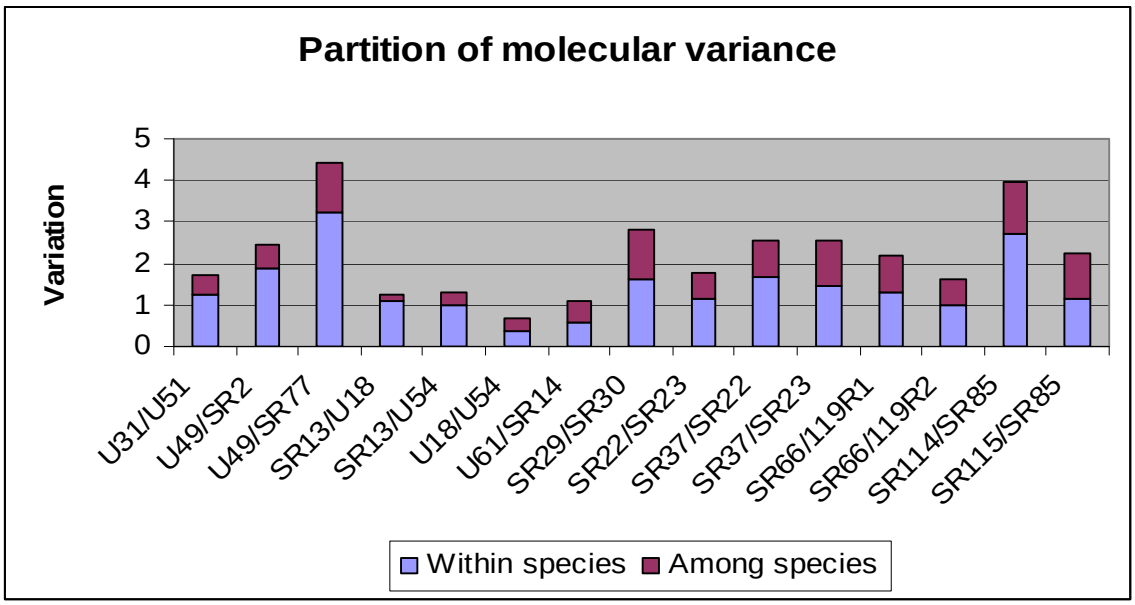
In summary, an examination of NJ trees and PCO plots showed that fragment profiles generated by almost all primer combinations were able to separate *S. aethnensis*, *S. chrysanthemifolius* and *S. aethnensis* from *S. vulgaris* and *S. cambrensis*. The latter two taxa formed a distinctive cluster using the datasets produced by primer pairs U31/U51, U49/snoR77, U61/snoR14, snoR29/snoR30, snoR37/snoR22, snoR22/snoR23, snoR37/snoR23, snoR66/119R1, snoR114/snoR85 and snoR115/snoR85. The primer combinations U18/U54 and snoR66/119R2 grouped only *S. vulgaris*, whereas *S. cambrensis* was found in a cluster using the snoR13/U54 primer pair. More distantly related species like *S. madagascariensis* and *S. flavus* tended to form separate groups in most analyses, and weak separation of the closely related species *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* was obtained using the primer combinations U49/snoR2, U49/snoR77, snoR37/snoR22, snoR37/snoR23, snoR66/119R1, snoR66/119R2 and snoR115/snoR85.

4.3.2.5 Analyses of molecular variance (AMOVA)

Because fragment profiles were not generated for all samples per primer pair, preliminary analyses of molecular variance (AMOVAs) were conducted on a subset of 43 samples that included *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis* across eight primer combinations U31/U51, U49/snoR2, snoR13/U18, U18/U54, U61/snoR14, snoR29/snoR30, snoR37/snoR22 and snoR22/snoR23. The results obtained from these analyses were similar to those obtained from AMOVAs performed on the full sample set. Thus, AMOVA was extended to all primer combinations that generated fragment length variation (Figure 4.9). Although analyses were conducted for different datasets (with and without *S. madagascariensis* included) and also on different numbers of 'species groups', only the AMOVA results for one dataset (without *S. madagascariensis*) that exclude 'species' groupings are shown (for all other results see supplemental material).

The amount of variation varied greatly between the different primer combinations, ranging from 0.664 (U18/U54) to 4.405 (U49/snoR77) (Figure 4.9). The percentage of total variation attributed to within species variation ranged from 52% (U61/snoR14 and

snoR115/snoR85) to 87% (SR13/U18). For all primer combinations except one (SR13/U18) variation among species was significant.



Cluster	Primer pair	No of Samples	Variation			Cluster	Primer pair	No of Samples	Variation		
			Total	Within species (%)	Among species (%)				Total	Within species (%)	Among species (%)
A	U31/U51	74	1.716	73	27	J	SR22/SR23	80	1.763	65	35
D	U49/SR2	52	2.457	76	24		SR37/SR22	57	2.572	65	35
	U49/SR77	38	4.405	73	27		SR37/SR23	38	2.528	57	43
E	SR13/U18	70	1.230	87	13	M	SR66/119R1	28	2.185	59	41
	SR13/U54	24	1.328	75	25		SR66/119R2	35	1.598	63	37
	U18/U54	55	0.664	56	44	N	SR114/SR85	24	3.950	68	32
F	U61/SR14	74	1.093	52	48		SR115/SR85	24	2.250	52	48
G	SR29/SR30	70	2.825	58	42						

Figure 4.9: Results of AMOVAs showing total variation, and percentage partitioned within and among species for 15 different primer combinations. P-values were estimated by 999 random permutations (not shown) and all variance components were highly significant ($p < 0.001$) for all primer pairs except SR13/U18. Analyses were conducted on datasets containing results for *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis* samples. (SR = snoR).

4.3.2.6 STRUCTURE assignment tests

STRUCTURE analyses were performed on all variable primer pair datasets and their subsets and the numbers of groups were chosen using the R-script STRUCTURE-SUM-2009.R. An example for choosing groups using "Structure.deltaK", "Structure.Table", "Structure.simil" plots and also various barplots are shown for variation of the snoRNA cluster G generated by the primer pair snoR29/snoR30 (Figure 4.10).

For the determination of the number of groups following the method of Evanno *et al*. (2005), four plots (Figure 4.10A-D) were generated by the "Structure.deltaK" function. The deltaK plot (Figure 4.10D) showed a high modal value at K=2 (black arrow) suggesting the presence of two groups in the dataset. However, a second peak, although much lower, was obtained for K=4 (red arrow) which might indicate the presence of four groups. While the plots of Mean L'(K) (Figure 4.10B) and Mean L''(K) (Figure 4.10C) reflect the findings of the Mean deltaK plot, the Mean L (K) plot (Figure 4.10) would suggest the presence of four groups due to the high likelihood value at K=4.

Similarly, it was concluded from the analysis of the L (K) plot generated by the "Structure.Table" function that there are four groups present in the dataset because K=4 has the highest LnP value and the curve starts to flatten out above this value which is indicated by the red arrow (Figure 4.10E). Furthermore, the replicates (5 runs for each K) show very high similarity as indicated by the similarity coefficient 1 shown in the "Structure.simil" plot (Figure 4.10F) and no empty group (groups with no sample assigned) are obtained (not shown). Because it was not possible to clearly determine the number of groups represented by the dataset three barplots (Figure 4.10G) were produced showing the assignment of the different samples into the two, three and four groups. For K=2 all samples of *S. vulgaris* (indicated by the green vertical bar), *S. cambrensis* (indicated by the yellow vertical bar) and *S. madagascariensis* (brown vertical bar) were assigned to one group (indicated by the dark red horizontal bands) and the samples of *S. aethnensis* (indicated by the red vertical bar), *S. chrysanthemifolius* (blue vertical bar) and *S. squalidus* (light purple vertical bar) were assigned to a third group (blue horizontal bands) to another group. While the barplot for K=3 splits off *S. madagascariensis* (yellow horizontal bands), the barplot for K=4 suggests an extra group consisting of some of the *S. squalidus* samples (indicated by turquoise horizontal bands). Some individuals

in the barplot for K=4 show mixed profiles which would assign them into different groups and in some cases indicate a hybrid origin. However, mixed individuals can be allocated to certain groups according to their corresponding bar length (in %). For example, some *S. squalidus* samples have bars belonging almost entirely to one group (indicated by turquoise horizontal bands, barplot K=4) and can with certainty be assigned to this group. Others like the *S. madagascariensis* sample mentioned above have fragment profiles with weaker "affinity" to one group and should probably not be assigned to any particular group (green arrow; Figure 4.10G). Interestingly, although the generated plots (Figure 4.10A-F) did not indicate the presence of three groups in the dataset, the barplot obtained for K=3 reflected the distant relationship between *S. madagascariensis*, *S. vulgaris/S. cambrensis* and *S. aethnensis/S. chrysanthemifolius/S. squalidus*. However, as shown in the plots it appears to be more likely that the real number of groups is four indicating some degree of substructure within the diploid species *S. squalidus*. It should be noted that other datasets did not result in plots as clearly differentiated as the one seen in Figure 4.10.

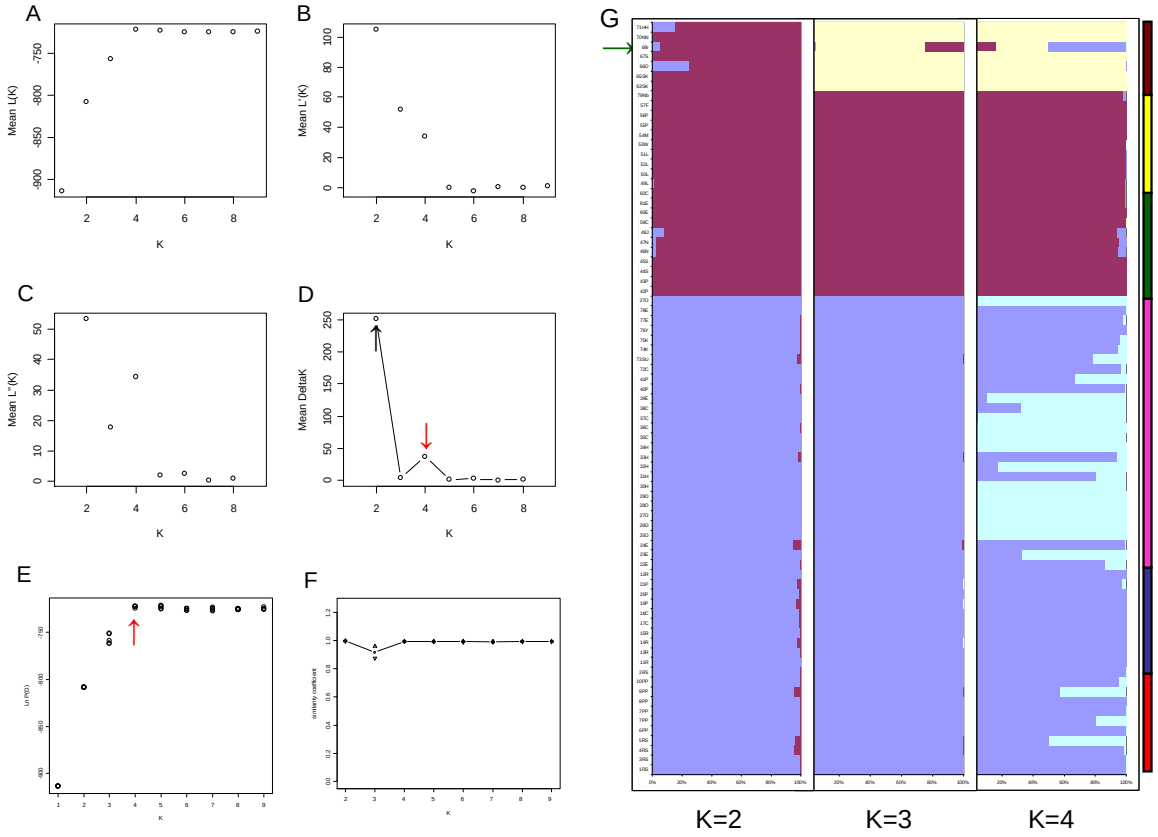


Figure 4.10: STRUCTURE analysis for the dataset of the snoR29/snoR30 primer pair. The "Structure.deltaK" function generates 4 plots: A: Mean L(K), B: Mean L'(K), C: Mean L''(K) and D: Mean deltaK. E: The "Structure.Table" plot shows the likelihood of data (five iterations) for increasing numbers of K. F: "Structure.simil" plot summarizes the similarities of the five iterations for each K. G: Barplots for the two, three and four groups (K=2, 3, 4) produced from the structure analysis with the highest likelihood value. Vertical bars represent different species: red = *S. aethnensis*, blue = *S. chrysanthemifolius*, light purple = *S. squalidus*, green = *S. vulgaris*, yellow = *S. cambrensis*, brown = *S. madagascariensis*. Green arrow = *S. madagascariensis* sample which was assigned to three different groups in the K=4 plot.

The number of groups obtained by the STRUCTURE analyses of datasets generated by all primer pairs and snoRNA clusters ranged from 1 to 4, and it was evident that the groups identified were relatively consistent across gene clusters (Table 4.8). Thus, *Senecio vulgaris* and *S. cambrensis* were grouped together by each of 6 primer pairs,

whereas *S. madagascariensis* samples, when included in analyses, were separated from these taxa by each of four primer combinations (Table 4.8 and supplemental material for more details).

Table 4.8: Numbers of groups resolved by each primer pair using Structure. *S. cam* datasets contain *S. aethnensis, S. chrysanthemifolius, S. squalidus, S. vulgaris* and *S. cambrensis* samples, whereas *S. madagascariensis* was included in the *S. mada* datasets. V = *S. vulgaris,* C = *S. cambrensis,* CR = closely related species, M = *S. madagascariensis,* Sq = *S. squalidus,* A = *S. aethnensis,* ng = no grouping. SR = snoR.

Cluster	Primer pair	Dataset	No of samples	Number of groups (K)	Species groups	Cluster	Primer pair	Dataset	No of samples	Number of groups (K)	Species groups
A	U31/U51	S. mada	83	4	V/C, M	J	SR22/SR23	S. mada	89	2	V/C
D	U49/SR2	S. cam	52	2	ng		SR37/SR22	S. cam	57	3	V/C
	U49/SR77	S. cam	38	2	A		SR37/SR23	S. cam	38	3	C, V
E	SR13/U18	S. mada	79	2	M	M	SR66/119R1	S. cam	28	2	V/C
	SR13/U54	S. cam	24	1	ng		SR66/119R2	S. mada	39	2	V/M
	U18/U54	S. mada	64	2	C/M	N	SR114/SR85	S. cam	24	1	ng
F	U61/SR14	S. mada	83	3	V/C, M		SR115/SR85	S. cam	24	2	V
G	SR29/SR30	S. mada	77	4	V/C, M, Sq						

4.3.2.7 Summary

The primer combinations differed in the number of fragment profile they generated across the samples examined and in their abilities to separate certain groups of samples (Table 4.9). For some primer pairs the same groupings could be obtained by both genetic distance (NJ and PCO analyses) and fragment frequency (STRUCTURE) based methods. For others, these two methods showed differences in the number and/or composition of groups detected with a tendency for fewer groups being detected using STRUCTURE (Table 4.9). For instance, the same groupings of samples were obtained by both methods from fragments generated by the primer pairs U31/U51, snoR13/U18, U61/snoR14,

snoR29/snoR30, whereas only the *S. vulgaris/S. cambrensis* (V/C) group was detected by both methods using snoR22/snoR23, snoR37/snoR22 and snoR66/119R1, and different groupings were produced using U49/snoR77 (V/C, *S. madagascariensis* (M) vs *S. aethnensis* (A)), U18/U54 (V, M vs C/M), snoR37/snoR23 (V/C, closely related (CR) vs C, V), snoR66/119R2 (V, M, CR vs V/M) and snoR115/snoR85 (V/C, CR vs V). The remaining primer combinations did not show any specific groupings of species (no grouping (ng)) which appears to coincide with the small sample size of all but one of these datasets, but may well merely reflect the relative low level of among species variation they contained. Nevertheless, most primers resolved a certain degree of structure, with most of them grouping *S. vulgaris* with *S. cambrensis,* distinguishing *S. madagascariensis,* and indicating some weak separation between the remaining closely related species examined.

Some of the fragments amplified were shared between hybrid taxa and their parents and might be used to detect hybridisation. In fact all primer combinations except U49/snoR77, snoR13/U18 and snoR13/U54 generated such fragments.

Table 4.9: Summary of the different analytical methods used to identify groups among samples of species investigated according to fragment profiles generated by different primer pairs. SR = snoR; V = *S. vulgaris*, C = *S. cambrensis*, CR = closely related species, M = *S. madagascariensis*, Ch = *S. chrysanthemifolius*, Sq = *S. squalidus*, A = *S. aethnensis*, ng = no grouping. AFF = analysis of fragment frequencies. Note that the first letter refers to the hybrid and the second one to the parent in the hybrid-parent fragments.

Cluster	Primer pair	No of samples genotyped	NJ, PCO	STRUCTURE	AMOVA	AFF
			Groupings	Groupings	% among species variation	Hybrid-Parent fragments*
A	U31/U51	85	V/C, M	V/C, M	27	C-V
D	U49/SR2	52	CR	ng	24	C-V, C-S
	U49/SR77	38	V/C, M	A	27	-
E	SR13/U18	88	M	M	13	-
	SR13/U54	24	C, Ch	ng	25	-
	U18/U54	69	V, M	C/M	44	C-V, C-S^
F	U61/SR14	88	V/C, M	V/C, M	48	C-V^
G	SR29/SR30	84	V/C, M, Sq	V/C, M, Sq	42	C-V, C-S
J	SR22/SR23	96	V/C, M	V/C	35	C-V, C-S, S-Ch^
	SR37/SR22	58	V/C, CR	V/C	35	C-V, C-S
	SR37/SR23	38	V/C, CR	C, V	43	C-S
M	SR66/119R1	30	V/C, CR	V/C	41	C-V, C-S, S-Ch^
	SR66/119R2	38	V, M, CR	V/M	37	C-S
N	SR114/SR85	24	V/C	ng	32	C-V, S-Ch^, S-A^
	SR115/SR85	24	V/C, CR	V	48	C-V, C-S, S-Ch^

* Only fragments which show high frequencies in the hybrid and one parent and are absent (0 to 5 % frequency) or present in low (< 30 %) frequencies (^) in the other parent.

Some of the primer pairs were used to amplify only a few samples (less than eight samples per species) and the results for these primers should be viewed cautiously. These primer combinations were not included in the analysis of combined data sets described below.

4.3.2.8 Combined datasets

Datasets for the snoRNA gene clusters A, D, E, F, G and J were combined and analysed. This was done on datasets after either removal of samples that were not profiled for each primer pair used (i.e. pruned dataset – P) or after inserting missing data (i.e. MD dataset). The datasets were chosen for combination based on the number of available fragment profiles; thus, the combined datasets consisted of 8 (U31/U51, U49/snoR2, snoR13/U18, U18/U54, U61/snoR14, snoR29/snoR30, snoR37/snoR22 and snoR22/snoR23), 7 (8 without snoR37/snoR22), 6 (7 without U49/snoR2), 5 (6 without snoR13/U18 and U18/U54, respectively) and 3 (snoR22/snoR23, snoR29/snoR30 and U31/U51) different primer pair data matrices. Most data matrices contained samples of *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris*, *S. cambrensis*, and *S. madagascariensis*; however, the matrices that combined data for eight and seven primer combinations, respectively, lacked samples of *S. madagascariensis* due to inclusion of data generated by the snoR37/snoR22 and U49/snoR2 primer pairs. One *S. teneriffae* sample was also examined for all 8 primer pairs and included in individual based analyses.

Very similar results (NJ, PCO and STRUCTURE analyses) were obtained for all combined datasets (see supplemental material for NJ trees). In general, three distinct clusters comprising *S. aethnensis*/*S. chrysanthemifolius*/*S. squalidus*, *S. vulgaris*/*S. cambrensis*/*S. teneriffae* and, when included in analyses, *S. madagascariensis* were obtained. Within these clusters species were partially separated from each other, albeit with various degrees of overlap. Samples of *S. vulgaris* and *S. cambrensis/S. teneriffae*, in particular, tended to be separated into different groups, whereas those of *S. aethnensis*, *S. chrysamthemifolius* and *S. squalidus* were more intermingled. As an example, the results for the dataset containing 6 primer pair matrices and missing data (6-MD) are illustrated in Figure 4.11A and B.

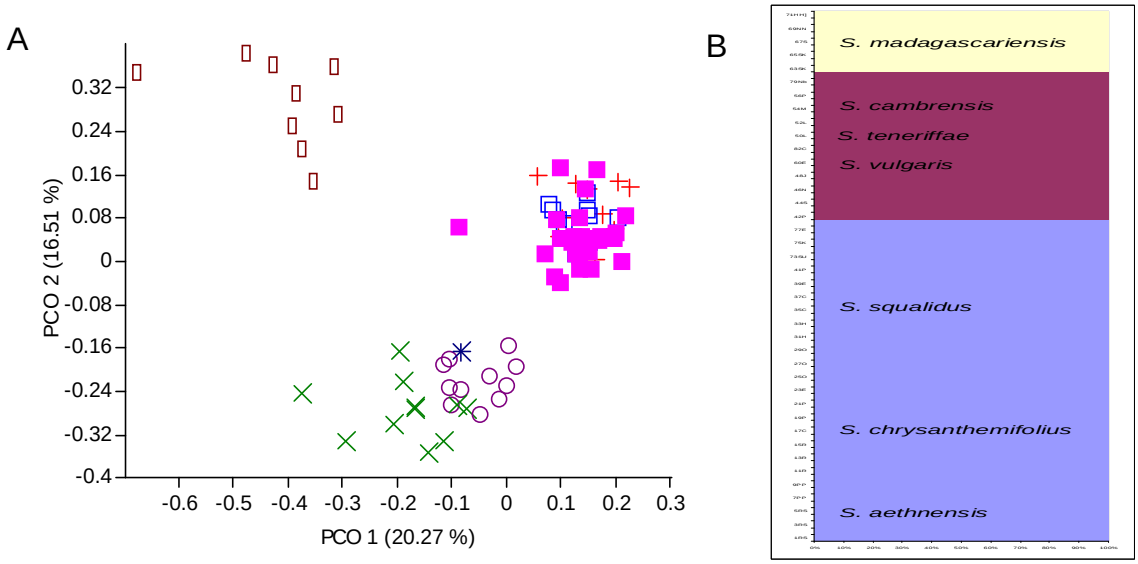


Figure 4.11: PCO (A) and STRUCTURE (B) plots from the analysis of combined MD datasets of six primer pairs. Three different groups are evident: one comprising *S. aethnensis*/*S. chrysanthemifolius*/*S. squalidus* (blue in B), one comprising *S. vulgaris*/*S. cambrensis*/*S. teneriffae* (dark purple in B), and a third comprising *S. madagascariensis* samples (yellow in B). *S. aethnensis* = +; *S. chrysanthemifolius* = □; *S. squalidus* = ■; *S. vulgaris* = **x**; *S. cambrensis* = ○.

A NJ tree produced from the analysis of the combined dataset of 8 primer pairs (excluding *S. madagascariensis*, Figure 4.12) shows a broadly equivalent pattern of relationships to that evident in the PCO plot (Figure 4.11A). *S. vulgaris*, *S. cambrensis* and one sample of *S. teneriffae* formed a clade having reasonably high bootstrap support (84 %), with two *S. vulgaris* samples (from Egypt (EGY) and Cardiff (CAR)) placed within *S. cambrensis*. Interestingly, all *S. cambrensis* samples from Edinburgh (EDI), which represent an independently originated lineage of the species, were grouped together (61 % bootstrap support). Most samples of *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* were grouped according to species, but without apparent geographical structure. However, five individuals of *S. squalidus* were intermingled with *S. aethensis* and one sample of each of *S. aethnensis* and *S. squalidus* were intermixed with *S. chrysanthemifolius* (Figure 4.12).

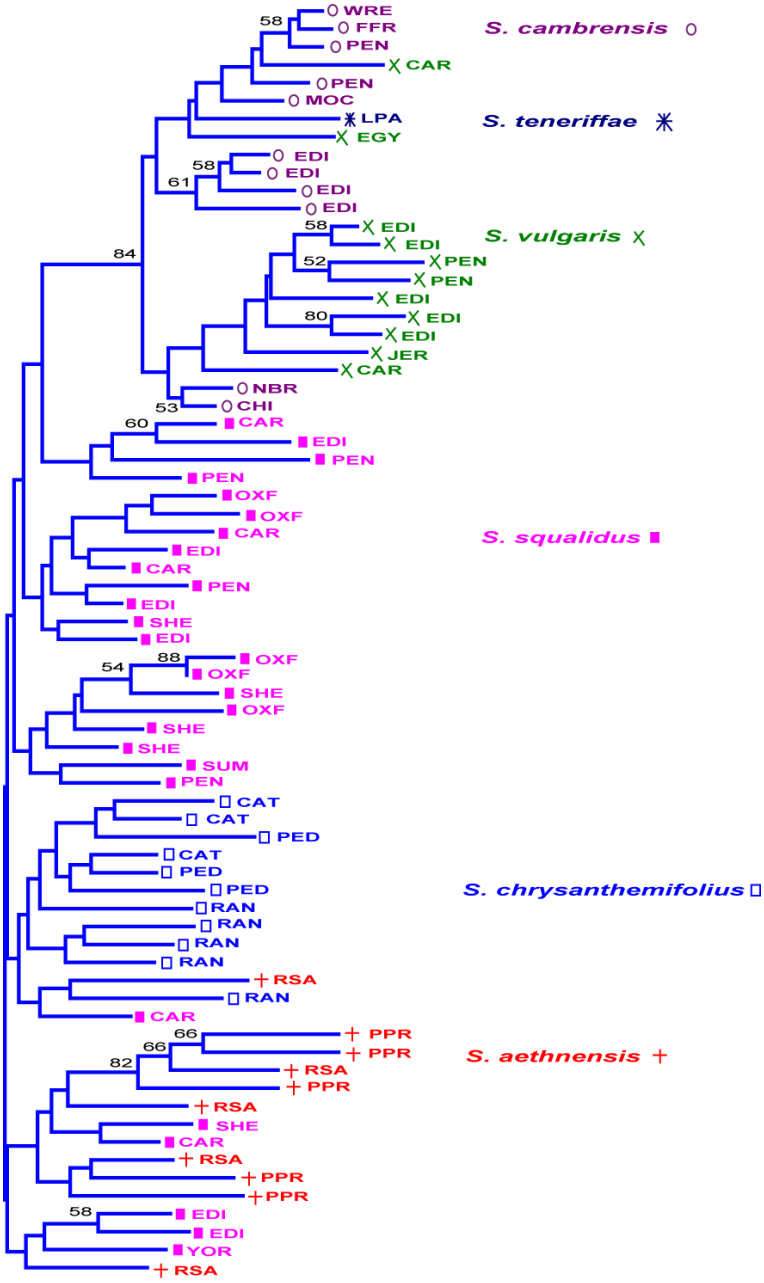


Figure 4.12: Relationships among 71 individuals of six *Senecio* species based on NJ analysis of fragment variation and dice similarity indices. Numbers above/below branches indicate bootstrap values (>50 %) generated from 1000 replicates. Locations of samples are shown next to branches. WRE = Wrexham, FFR = Frith, PEN = Pentre, CAR = Cardiff, MOC = Mochdre, LPA = La Palma, EGY = Egypt, EDI = Edinburgh, JER = Jersey, NBR = New Brighton, CHI = Chirk, OXF = Oxford, SHE = St. Helens, SUM = Summerhill, YOR = York, CAT = Catania, PED = Pedara, RAN = Randazzo, RSA = Rifugio Sapienza, PPR = Piano Provenziana.

An AMOVA (Table 4.10) conducted over all five species showed that 63% of the fragment diversity was due to variation between individuals within species and 37% was explained by differences between species. Separate AMOVAs conducted on each of the two groups identified by NJ, PCO and STRUCTURE analyses showed that 22% of total variation was accounted for by differences among *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus*, and 19 % for differences between *S. vulgaris* and *S. cambrensis*. Similar results were obtained for all other combined datasets with among species variation for all five of these species ranging from 60 to 67%.

Table 4.10: Analysis of molecular variance (AMOVA) conducted on fragment variation of the combined dataset for eight primer combinations among and within *S. aethnensis, S. chrysanthemifolius, S. squalidus, S. vulgaris* and *S. cambrensis*. P-values were estimated by 999 random permutations.

Source of variation	d.f.	Sum of squares	Estimated variance	Percentage of variation	P
S. aethensis/S. chrysanthemifolius/S. squalidus - S. vulgaris/S. cambrensis					
Among species	4	205.883	3.435	37%	<0.001
Within species	65	388.376	5.975	63%	<0.001
S. aethensis/S. chrysanthemifolius/S. squalidus					
Among species	2	96.926	2.743	22%	<0.001
Within species	45	439.704	9.771	78%	<0.001
S. vulgaris/S. cambrensis					
Among species	1	29.794	1.967	19%	<0.001
Within species	20	163.205	8.160	81%	<0.001

All pairwise Φ_{ST} values between species (Table 4.11) were significant and indicate that *S. chrysanthemifolius* is more similar to both *S. squalidus* (Φ_{ST} = 0.188) and *S. aethnensis* (Φ_{ST} = 0.223) than these two species are to each other (Φ_{ST} = 0.249). *S. vulgaris* and *S. cambrensis* show a similar amount of differentiation (Φ_{ST} = 0.228) but differed greatly from each of the other species.

Table 4.11: Mean pairwise genetic differentiation (Φ_{ST}) between S. aethnensis, S. chrysanthemifolius, S. squalidus, S. vulgaris and S. cambrensis. P-values were estimated by 999 random permutations and all Φ_{ST} values are highly significant ($p < 0.001$).

	S. aethnensis	S. chrysanthemifolius	S. squalidus	S. vulgaris	S. cambrensis
S. aethnensis	0.000				
S. chrysanthemifolius	0.223	0.000			
S. squalidus	0.249	0.188	0.000		
S. vulgaris	0.409	0.494	0.449	0.000	
S. cambrensis	0.381	0.466	0.387	0.228	0.000

Separate AMOVAs conducted for each species resulted in the following percentage of variation between populations: *S. aethnensis* (7 %), *S. chrysanthemifolius* (10 %), *S. squalidus* (14 %), *S. vulgaris* (9 %) and *S. cambrensis* (40 %). However, only the results for *S. squalidus* and *S. cambrensis* were highly significant ($p = 0.001$ and 0.003, respectively).

4.3.2.8.1 Ancestry of hybrid species *S. squalidus* and *S. cambrensis*

All but three samples (two of *S. aethnensis* and one of *S. chrysanthemifolius*) showed zero probability of being derived from the cluster of the other parent species and, thus, can be regarded as pure representatives of parent species. Estimates of the ancestry of 27 samples of *S. squalidus* and 11 individuals of *S. cambrensis* were obtained using the MD combined dataset of all 8 primer combinations (Figure 4.13). In *S. squalidus* the proportion of ancestry derived from *S. chrysanthemifolius* ranged from 11.2 to 96.8 % (mean 58.2 %, SD = 33.6). Interestingly, all samples from Oxford showed a high proportion of *S. aethnensis* ancestry, whereas all but one individual from Edinburgh exhibited a very high level of *S. chrysanthemifolius* ancestry (Figure 4.13A). All samples of *S. cambrensis* analysed showed a low amount of mixed ancestry with a maximum of 13.4 % derived from its *S. squalidus* parent (Figure 4.13B).

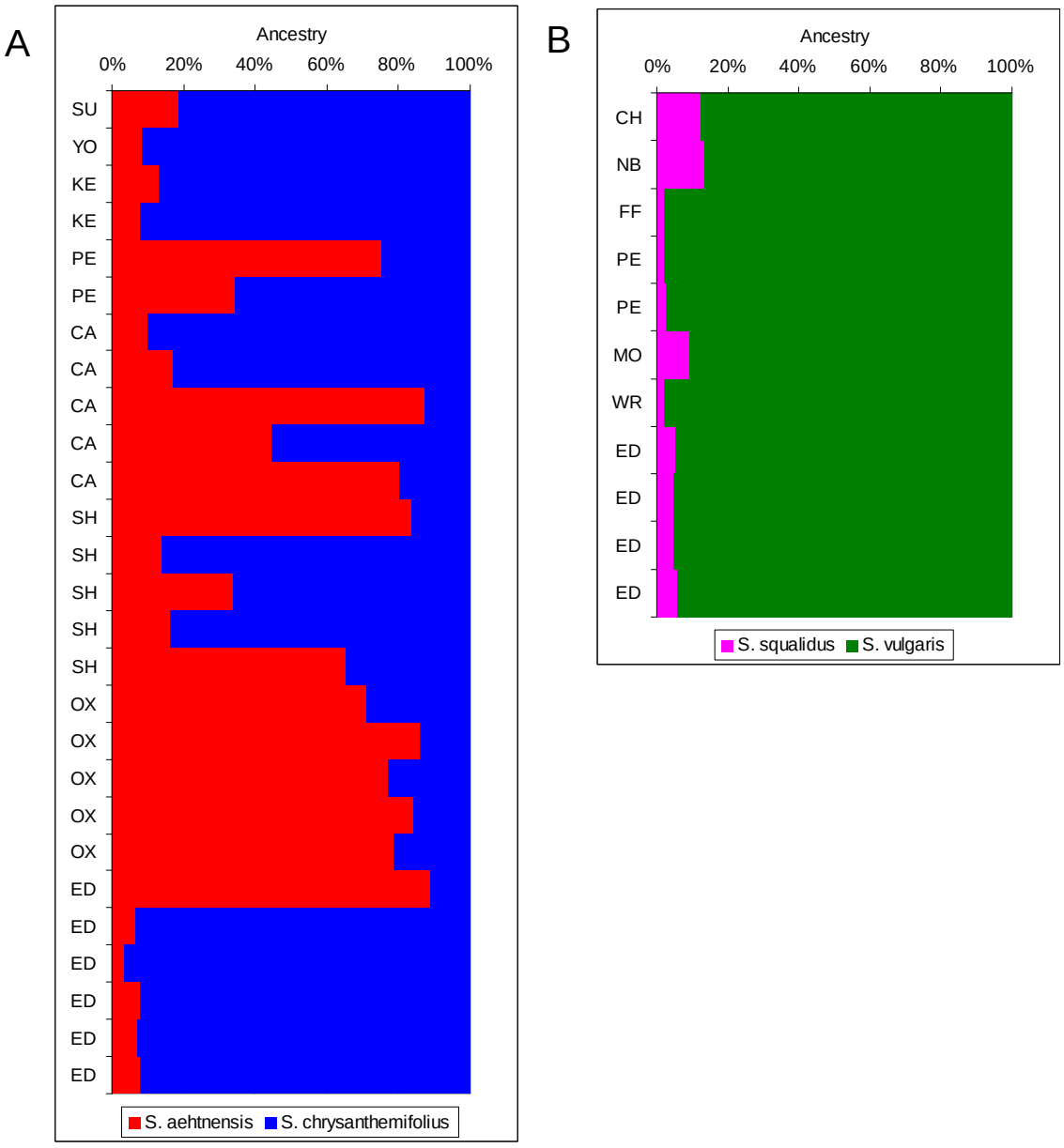


Figure 4.13: Proportion of parents' ancestry in the hybrid species *S. squalidus* (A) and *S. cambrensis* (B). A: The 27 *S. squalidus* individuals analysed were sampled from eight locations across Great Britain. B: The 11 samples of *S. cambrensis* analysed were collected from seven British populations. ED = Edinburgh, OX = Oxford, SH = St. Helens, CA = Cardiff, PE = Pentre, KE = Kent, YO = York, SU = Summerhill, WR = Wrexham, MO = Mochdre, FF = Ffrith, NB = New Broughton, CH = Chirk.

The species, *S. flavus*, *S. engleranus*, *S. glaucus*, *S. massaicus*, plus two additional samples of *S. teneriffae*, were only profiled using some primer pairs. An analysis including these samples produced a NJ tree (see appendix) similar to the NJ tree of the combined dataset in the initial investigation (see Figure 4.5). While *S. flavus*, *S. engleranus*, *S. massaicus* were placed in a clade with *S. madagascariensis*, *S. glaucus* was positioned closer to *S. aethnensis*. The two extra *S. teneriffae* samples were placed within *S. squalidus* and at the base of the *S. vulgaris*/*S. cambrensis* clade.

4.3.2.9 Genetic distance between NJ trees

The tree topologies generated from all pruned single primer pair data matrices used in the combined dataset analyses, together with the tree for the combined dataset (8-P dataset) were compared using TREEDIST implemented in the PHYLIP package. Because the U18/U54 NJ tree differed greatly from all of the other NJ trees (not shown) it was removed from analysis. The PCO plot (Figure 4.14) obtained from the distances between the remaining trees placed the snoR37/snoR22 (■) and U49/snoR2 (▲) trees closest to that for the combined dataset (blue ♦). Other trees were more divergent in topology from these three trees.

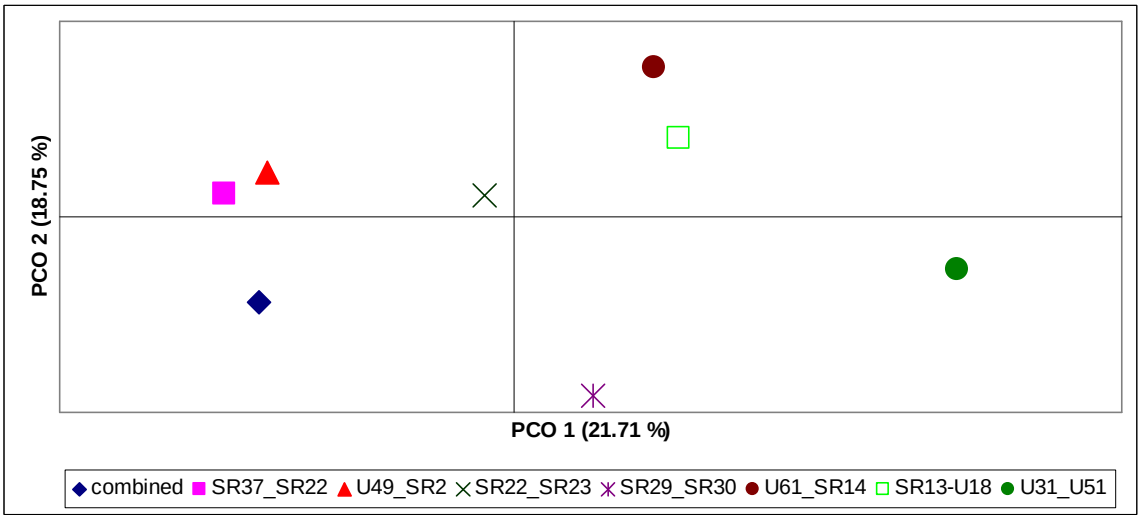


Figure 4.14: Tree distance analysis of various NJ trees. The PCO plot shows the distances between the different NJ tree topologies of seven single datasets (without U18/U54) and the resulting combined 8-P dataset (combined). The first two axes of the PCO plot explain 40.46 % of the total variation. Note that each NJ tree used for calculation consisted of the same 43 samples comprising *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis*.

4.4 Discussion

The aim of the work reported in this chapter was to determine whether primer combinations designed to amplify snoRNA genes and clusters in *Arabidopsis thaliana* might be used to examine patterns of genetic variation within and among *Senecio* species. The primer combinations were tested for amplification and profiled using radioactive and/or fluorescence labelling. Differences in amplification success between these two techniques are likely to be due to different PCR protocols optimized for each of them. An initial investigation across a wide range of *Senecio* species showed that amplification using radioactively labelled primers resolved fragment length variation for most primer pairs employed. Following this, a survey of snoRNA variation within and between species was conducted using fluorescence labelled primers on a higher number of samples of mainly *S. aethnensis*, *S. chrysanthemifolius* and their homoploid hybrid species *S. squalidus*, plus *S. vulgaris* and the allopolyploid *S. cambrensis*, which is derived from hybridization between *S. vulgaris* and *S. squalidus*. Also included in the second analysis for many but not all primer combinations were several samples of *S. madagascariensis*.

4.4.1 Universality and simplicity

One criterion for molecular markers used in phylogenetics and DNA barcoding is universality (Kress *et al.*, 2005; Savolainen *et al.*, 2005) meaning that the regions examined are present across a wide range of species (e.g. land plants) and can be amplified using the same sets of primers (Alvarez & Wendel, 2003; Hollingsworth *et al.*, 2009b). Currently most universal markers available are specific to the chloroplast genome and the nuclear ribosomal DNA (Small *et al.*, 2004). They are present in a high number of copies and are, therefore, relatively easy to amplify using a standardised protocol thus adding simplicity to the procedure. More experimental expertise is assumed to be required for most single and low copy nuclear regions (Alvarez & Wendel, 2003). All primer pairs used in this study were designed from *A. thaliana* single and low copy snoRNA gene/gene cluster sequences. Most of these sequences appear to fulfil both

universality and simplicity criteria because they were successfully amplified in either all or the majority of *Senecio* species using a standardized procedure. Only two primer pairs, both of which included the snoR4 primer, did not amplify using either the radioactive or fluorescence labelling techniques. This failure may have been caused by the mismatch of the primer to annealing sites.

Some samples, in particular *S. squalidus* individuals, could not be amplified for every primer pair. PCR amplification problems with this species have been reported by others (A. Brennan and D. Forbes at St. Andrews University, and M. Hegarty at Aberystwyth University, personal communications), and it is feasible that the PCR reaction was inhibited by substances within the DNA extracts of these plants (Gagneux *et al.*, 1997). Furthermore, the fragment profiles generated by most primer pairs using a standardized protocol differed considerably in fragment numbers generated per sample. While these differences can be partly explained by allele number (heterozygotes vs. homozygotes), some additional fragments generated might be nonspecific amplification products (Kohn & Wayne, 1997) and/or artifacts of the scoring procedure (e.g. scoring threshold (Bonin *et al.*, 2004). The primers used were very specific, but they might have annealed to and amplified non-snoRNA gene/gene cluster regions with similar primer sequences in low stringency PCRs. However, these nonspecific fragments might be amplified from very specific loci and, therefore, provide additional phylogenetic information.

Contamination and the quality and quantity of DNA extracts (Taberlet *et al.*, 1996; Matsuzaki *et al.*, 2004) might also contribute to variation in fragment number. Thus, all DNA extracts were checked prior to amplification and poor quality fragment profiles were removed from datasets. Although the contribution of all of these factors to variation cannot be completely ruled out, it is likely they account for only a minor part of the total variation detected. The error rate calculated for the U14-3/U14-4 dataset based on replicate samples was 10 %, and 7 % could be attributed to errors in scoring double bands which might be the result of biochemical factors. This error rate is intermediate to rates usually obtained for RAPDs and ISSRs (15 to 25 %) and for RFLPs, SSRs and AFLPs (2 to 5%) (McGregor *et al.*, 2000; Bonin *et al.*, 2004; Koopman, 2005; Pompanon *et al.*, 2005; Meudt & Clarke, 2007; Cárdenas-Flores *et al.*, 2010).

4.4.2 SnoRNA gene/gene cluster variation between and within *Senecio* species

Analyses of most datasets for each primer pair separated distantly related species and sometimes more closely related species from each other, according to snoRNA fragment variation exhibited. However, most single pair primer datasets contained relatively few fragments which can often cause the generation of unstable and star like trees based on individual variation (Hollingsworth & Ennos, 2004). The topology of such trees might change markedly by variation in a single fragment and, thus, errors could greatly affect interpretation. Therefore, single primer pair datasets were combined for further analysis to improve the resolution of species relationships based on variation across several different genomic regions. Thus, the following discussion is based on the results from the analysis of combined datasets.

The initial primer-trial analysis, which employed radioactively labelled primer pairs, showed that samples of *S. vulgaris* did not form a distinct clade, but instead were grouped with samples of *S. aethnensis, S. chrysanthemifolius, S. squalidus, S. cambrensis* and *S. teneriffae*. Although some substructure was obtained within this group of taxa, species were often intermixed. The subsequent more detailed analysis of snoRNA variation, employing fluorescence labelled primers, focussed particularly on relationships between species of this group and surveyed a much higher number of samples of *S. aethnensis, S. chrysanthemifolius, S. squalidus, S. vulgaris* and *S. cambrensis,* and also of *S. madagascariensis*. The application of fluorescence labelled primers increased the sensitivity of fragment analysis by yielding a higher number of fragments which most likely accounted for the better resolution obtained. Analysis of the data from this analysis of combined data sets tended to separate species into three distinct groups - *S. madagascariensis, S. vulgaris*/*S. cambrensis,* and *S. aethnensis*/*S. chrysanthemifolius*/*S. squalidus*. The difference between the latter two groups of taxa was reflected by high bootstrap support of the *S. vulgaris*/*S. cambrensis* group and high Φ_{ST} values between them. Within these two groups, species were separated from each other, albeit with some degree of overlap. In particular, some *S. vulgaris* samples were intermingled with *S. cambrensis* as was also found to be the case in a previous study of AFLP variation within and among these taxa (Abbott *et al.*, 2007). Pairwise Φ_{ST} values between species

indicated that *S. chrysanthemifolius* was genetically more similar to *S. squalidus* than to *S. aethnensis*.

In contrast to the situation in the other species examined, a high amount of variation (40%) was shown to be present within *S. cambrensis*, which reflects the fact that two different independently originated lineages (Welsh and the Edinburgh lineages) of this species (Abbott, 1992; Harris & Ingram, 1992a) were included in the survey.

In the NJ trees of the primer-trial analysis (Figure 4.5) *S. mohavensis* and *S. glaucus* were found close to *S. flavus* which appears not surprising as the former species originated by hybridisation of the latter two species (Liston & Kadereit, 1995; Comes & Abbott, 2001; Coleman *et al.*, 2003; Kadereit *et al.*, 2006). Interestingly, *S. flavus* and *S. glaucus* are very distantly related. In various ITS phylogenies *S. mohavensis* and *S. glaucus* were placed within a poorly resolved clade containing also *S. squalidus*, *S. chrysanthemifolius*, *S. aethnensis* and other species (i.e. *S. squalidus* clade (ITS phylogeny in Chapter 1); also called Mediterranean complex (Comes & Abbott, 2001), Groundsel clade III (Coleman *et al.*, 2003) and clade A (Pelser *et al.*, 2007)), while *S. flavus* was placed together with *S. engleranus* in a most distant position relative to this clade. These two species, *S. flavus* and *S. engleranus*, may not even be part of *Senecio sensu stricto*, but most closely related to the genus (Pelser *et al.*, 2007; Milton, 2009). Interestingly, RAPD analysis of 10 selected species of the Mediterranean complex, including *S. mohavensis*, *S. glaucus* and *S. flavus*, placed these three species within the same clade (clade A) and, thus, suggests a much closer relationship between these species (Comes & Abbott, 2001). Although a different set of species were used in the study presented here, the results were similar to the RAPD analysis and, therefore, support the findings of Comes & Abbott (2001).

S. glaucus was also found close to the group of species containing *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* in the NJ trees of the primer-trial analysis. The *S. glaucus* samples placed next to *S. flavus* were collected in Israel, whereas the one phylogenetically close to *S. squalidus* group was sampled in Morocco. Samples from these locations differed considerably in their ITS sequences, showed high variation in their cpDNA haplotype (Comes & Abbott, 2001) and intraspecific geographical structure was shown by alloenzyme data (Comes & Abbott, 1999). Therefore, the high variation

within *S. glaucus* found in the study presented here is in accordance with previous examinations and might be explained by strong geographical barriers and selection between different populations (Comes & Abbott, 1999).

Although samples of *S. flavus* were also collected from very distant sites, low within species variation was obtained. This was also evident in their ITS sequences and might be explained by relatively recently colonisation due to Pleistocene migration and/or long distance dispersal mediated by birds (Coleman *et al.*, 2003). Less intraspecific variation was obtained for *S. mohavensis,* but it was possible to distinguish between the two disjunct subspecies *mohavensis* and *breviflorus*, respectively. The lack of ITS sequence variation is thought to be a result of the recent origin and disjunction (mediated by long distance dispersal from southwest Asia to North America) of *S. mohavensis* ssp. *mohavensis* (Coleman *et al.*, 2003). Both, *S. flavus* and *S. mohavensis* are self-fertile which is a great advantage in long distance dispersal due to the possibility of single colony establishment, decreased inbreeding depression and low pollinator dependence. However, self-fertilisation can result in a reduced amount of genetic variation and might, therefore, contribute to the low within species variation observed.

Both morphological and isoenzyme data suggested a considerable amount of variation within the *S. madagascariensis* complex (Radford *et al.*, 2000) and ITS data showed some degree of intraspecific sequence differentiation (Le Roux *et al.*, 2006). Thus, it is not surprising that relatively high intraspecific variation was obtained for *S. madagascariensis* in this study in both the initial primer trial and the more detailed analysis.

4.4.3 Hybrid origin of various Senecio species

Part of the present study of snoRNA gene/gene cluster length variation was aimed to investigate whether such variation is of use in studying the hybrid origins of diploid *S. squalidus*, the hexaploid *S. cambrensis,* and the tetraploid *S. mohavensis* ssp. *breviflorus*.

S. mohavensis ssp. *breviflorus*, originally described as *S. flavus* ssp. *breviflorus*, originated from a cross between *S. glaucus* and *S. flavus* in which the former species acted as the female parent (Comes & Abbott, 2001). The ITS sequence of *S. mohavensis*

ssp. breviflorus was then homogenised towards *S. glaucus* (Comes & Abbott, 2001; Coleman *et al.*, 2003). Molecular evidence for the involvement of *S. flavus* in this cross was obtained from a survey of RAPD (Comes & Abbott, 2001) and AFLP variation (Kadereit *et al.*, 2006). In these studies, *S. flavus* ssp. *breviflorus* (i.e. *S. mohavensis* ssp. *breviflorus*) was found to be phylogenetically close to *S. flavus* rather than *S. glaucus*. Although only a few samples of *S. flavus, S. glaucus* and *S. mohavensis* were included in the study presented in this thesis very similar results were obtained.

S. squalidus is the homoploid hybrid of *S. aethnensis* and *S. chrysanthemifolius*. Morphologically, *S. squalidus* is an intermediate between its putative parents and, thus, indicates its hybrid origin (Abbott *et al.*, 2000; Abbott *et al.*, 2002). Initial molecular support for the hybrid origin of *S. squalidus* was provided by alloenzyme variation and later confirmed by RAPD/ISSR markers (Abbott *et al.*, 2002; James & Abbott, 2005). In the latter study 11 of 13 markers present in high frequency in *S. chrysanthemifolius* and absent or in low frequency in *S. aethnensis* and 10 of 13 markers for which the reverse was true were found in *S. squalidus*. However, only 11 and 7 markers, respectively, were shown to have high frequency in *S. squalidus*. In the study presented here, two fragments were found in high frequencies in *S. squalidus* and *S. aethnensis* and were absent or exhibited low frequency in *S. chrysanthemifolius* and six fragments were identified for which the reverse was true. The low number of diagnostic hybrid-parent fragments can be explained by the close genetic relationship between the parent taxa expressed by their many shared fragments. This is also shown by the study of RAPD/ISSR markers (James & Abbott, 2005) where only 65 of the 305 primer pairs (21 %) screened produced well-resolved bands which were able to distinguish between the two parent species. However, the higher number of snoRNA fragments shared between *S. squalidus* and *S. chryanthemifolius* would suggest a greater genomic proportion of *S. chrysanthemifolius* in *S. squalidus*. This is also supported by the estimate of ancestry from 27 individuals of *S. squalidus* based on snoRNA variation, which showed that mean proportion of the genome of *S. squalidus* derived from *S. chrysanthemifolius* was 58.2%. This value is similar to the 64.4 % based on RAPD/ISSR marker variation reported by James & Abbott (2005). However, the proportion of *S. chrysanthemifolius* ancestry based on snoRNA variation varied greatly between samples and was notably low in those from Oxford but

high in those from Edinburgh. This pattern might reflect the morphological variation within this species, with some individuals showing almost *S. chrysanthemifolius* phenotypes and others possessing more characteristics of *S. aethnensis* (James & Abbott, 2005).

The allohexaploid *S. cambrensis* originated by hybridisation between *S. vulgaris* and *S. squalidus*. Evidence for at least two independent origins of this species, one in Edinburgh, and another in North Wales, was obtained from surveys of isoenzyme and chloroplast variation ((Abbott, 1992; Ashton & Abbott, 1992; Harris & Ingram, 1992a). Morphologically, this species is mostly a mixture of traits possessed by either *S. squalidus* or *S. vulgaris* but some of its characters differ significantly from both parents and, thus it forms a distinctive morphological group (Abbott & Lowe, 2004). In an analysis of AFLP variation in samples of *S. cambrensis* and its two parent species collected from Wales (Abbott *et al.*, 2007), *S. cambrensis* was placed more closely to *S. vulgaris* with some degree of overlap in the PCO plot produced. In the study presented here, the estimate of ancestry using STRUCTURE showed a very high proportion of *S. vulgaris* in all 11 *S. cambrensis* individuals analysed. Furthermore, 32 fragments were shared in high frequencies between *S. cambrensis* and one parent and were absent or in low frequency in the other – 12 were shared with *S. squalidus* and 20 with *S. vulgaris*. The high number of shared fragments (relative to the shared fragments found in *S. squalidus* and its parents, *S. aethnensis* and *S. chrysanthemifolius*) might reflect the more distant relationship between the two parent species involved. The higher genetic proportion of *S. vulgaris* in *S. cambrensis* is thought to be caused by the two genomes inherited from *S. vulgaris* relative to one *S. squalidus* genome. Furthermore, although introgression cannot be ruled out, intergenomic recombination is suggested to be more likely to explain the higher similarity between *S. vulgaris* and *S. cambrensis* (Abbott *et al.*, 2007).

Morphologically very similar to *S. cambrensis* is *S. teneriffae*, another hexaploid hybrid species endemic to the Canary Islands, which most likely originated from a cross between *S. vulgaris* and *S. glaucus* (Lowe & Abbott, 1996; Abbott & Lowe, 2004). In the study presented here, one *S. teneriffae* individual was examined for all of the combined primer pairs in the more detailed analysis and showed much greater similarity to *S.*

vulgaris than to *S. glaucus*. This would suggest the same or similar mechanisms proposed for *S. cambrensis* such as the different numbers of genomes inherited (i.e. two genomes from *S. vulgaris* to one from *S. glaucus*), intergenomic recombination and, although not very likely, introgression. Interestingly, one of the extra *S. teneriffae* samples were placed closer to *S. squalidus* than to both *S. vulgaris* and *S. glaucus* which would suggest an involvement of *S. squalidus* rather than *S. glaucus* in the origin of *S. teneriffae*. However, the latter result is based on only two *S. teneriffae* samples and one individual of *S. glaucus* analysed for only a few primer pairs and in the initial investigation, where *S. vulgaris*, *S. teneriffae* and S. *squalidus* were intermixed, and should therefore be taken with caution. A more detailed survey using many more samples of both *S. glaucus* and *S. teneriffae* is required to investigate the origin of *S. teneriffae*.

4.4.4 Combining Datasets

As shown in the study presented here, some very useful patterns of variation emerged from the analysis of combined datasets making clear that surveys of snoRNA gene/gene cluster length variation are useful for examining phylogenetic relationships within closely related groups exhibiting some reticulate evolution. However, one crucial issue that arises when different single datasets are combined is that each of these datasets reflects a particular phylogenetic history which might or might not be similar to that reflected by other datasets (Tateno *et al.*, 1982). Incongruence between different datasets emerges through various processes (Meng & Kubatko, 2009). Furthermore, a subset of the produced datasets might already be enough to represent the relationship between the species. Additionally, different datasets might be used for different analysis. For example, some datasets might be more useful to investigate more distantly related species whereas others might be able to separate more closely related species. In this study, all variable datasets were subjected to various analyses (FFA, NJ, PCO, AMOVA, STRUCTURE) to explore their variability and the abilities to cluster certain groups, to provide diagnostic hybrid-parent and species specific fragments. Furthermore, the NJ trees of 8 single datasets (made up of 6 snoRNA gene clusters) were compared with each other and with the NJ tree of combined matrix using TREEDIST and showed that the combined dataset

is most similar to the snoR37/snoR22 dataset and most different to the U18/U54 matrix. The snoR37/snoR22 primer combination shows similar among species variation and is able to group most of *S. vulgaris*/*S. cambrensis* samples and separates *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus*, albeit with great overlap. Therefore, all these methods might be help in choosing regions for further investigation.

4.5 Conclusion

In this study, fragment length variation of an initial set of snoRNA genes/gene clusters was tested for their application in phylogenetic studies using a variety of *Senecio* species. All primer pairs were designed using *Arabidopsis thaliana* sequences and most of them showed amplification in the majority of species using a standardized protocol. The fragment profiles produced showed variation between and within species and by combining some of the datasets the results obtained were in accordance with previous studies mostly based on RAPD, AFLP and RAPD/ISSR markers in the delimitation of species and detection of reticulate evolution. Therefore, snoRNA gene/gene cluster fragment length polymorphisms (SRFLPs) can be used as a universal marker system for studying phylogenetic relationships between closely related species. However, to confirm that the amplification products are snoRNA genes/gene clusters these fragments should be sequenced. Sequencing would also provide information on the number of gene copies present, the sequence variation between orthologous and putative paralogous genes and might be used for isolating single copy regions which could then be used as codominant markers. Because snoRNA gene and genes/clusters are spread across the whole genome this marker system might also be used in the future for comparative mapping and to study the evolution of genes and genomes.

Chapter 5: SnoRNA genes and gene clusters in *Senecio*

5.1 Introduction

Although amplification success in *Senecio* of primer pair sequences based on snoRNA sequences in *Arabidopsis thaliana* suggests a similar gene cluster organisation in both genera, differences are possible in the number and size of fragments amplified. While the number of fragments amplified can be used to estimate the number of putative gene/gene cluster copies, size differences might reflect gene reorganisation within a cluster (e.g. differences in gene order, gene losses, as well as duplications and inversions). For example, a primer pair might produce a fragment in *Senecio* similar in size to a fragment expected in *A. thaliana* but in addition might amplify an extra and much longer fragment. This would suggest either two gene copies, one similar to that in *A. thaliana* and one with a long intergenic region, or a tandem repeat duplication of one gene, which is probably more likely. The major aim of the work reported in this chapter was to characterize snoRNA genes and gene clusters in *Senecio* species and determine differences in the organisation of snoRNA gene clusters relative to those in *A. thaliana*.

5.2 Material and Methods

SnoRNA gene clusters in *Senecio* were characterized by comparing the sizes of high and moderately frequent fragments, particularly from the diploid species *S. aethnensis, S. chrysanthemifolius* and *S. squalidus,* with fragments amplified by the same primers from *A. thaliana* and other species using ePCR. Blast searches based on *A. thaliana* snoRNA gene/gene cluster sequences were also performed (see Chapter 3). Various snoRNA genes plus primer sites within these sequences were identified, their organisation examined, and the sizes of possible PCR amplification fragments, together with sizes of genes and intergenic regions, were calculated. Most gene sizes should be relatively constant across species and, therefore, intergenic regions within *Senecio* were estimated by assuming gene sizes similar to other species, particularly those in *A. thaliana*. Some genes might show greater size variation and these were characterized using the gene size

of the most similar fragments. Overall, fragment length differences observed between species were assumed to be almost entirely intergenic or the result of differences in gene cluster organisation. It should be noted that most primer pairs should bind within two neighbouring genes (i.e. neighbours in *A. thaliana*) and, thus, should only amplify one intergenic region. Some gene clusters were examined by more than one primer pair, thus allowing a more reliable characterization of these clusters.

5.3 Results

BLAST searches based on *A. thaliana* snoRNA gene cluster sequences resulted in the identification of sequences from various species. The ESTs obtained may not be full length sequences with some lacking 5' and 3' ends. However, within these sequences, snoRNA genes, plus intergenic and primer sequences, were identified and their lengths were calculated. For snoRNA genes and gene clusters detected, see Tables and Figures in appendix, gene clusters M and N were previously shown as clusters D and E in Chapter 3. As expected, all but one of the snoRNA genes found in different species were relatively constant in size and most variation was due to intergenic size variation. The box C/D snoRNA gene U49, which is present in three copies in *A. thaliana*, differed considerably in size ranging from 75 bp in *Helianthus paradoxus* to 246 bp in *A. thaliana*. The organisation of most gene clusters appears to be strictly conserved with differences evident for only a few species examined. For example, in *Brassica oleraceae* the order of two adjacent genes of cluster A, i.e. snoR4 and U31, was inverted (see appendix, Figure A.16).

By comparing fragment sizes obtained from all primer pairs of clusters and assuming the gene sizes and organisation existing in *A. thaliana*, it is possible to estimate the size of intergenic regions which could accommodate additional genes. Furthermore, possible tandem repeats, gene losses and inversions might be identified. As an example, the reconstruction of gene cluster A is shown below. This cluster was chosen because three different primer pairs amplified it successfully, thus providing a particularly complete picture of it in *Senecio*.

5.3.1 Reconstruction of snoRNA cluster A in *Senecio*

The U33/U51 primer pair produced only one fragment of ca. 150 bp in *Senecio*, which was similar in length to fragments amplified by the same primers in *A. thaliana*. Therefore, the intergenic U33/U51 region is concluded to be approximately 40 bp long. Approximately 80 bp of the snoR4 gene is located between U31 and U33 in *A. thaliana*. Two fragments found in *Senecio* were similar in length to those in *Arabidopsis* detected by reverse ePCR. Their estimated U31/U33 intergenic regions of about 190 bp were sufficient to accommodate the snoR4 or any other snoRNA gene (in bold, Table 5.1; Appendix, Figure A.16).

In addition to the fragment likely to accommodate a putative snoRNA gene, fragments with U31/U33 intergenic regions of about 50 bp were obtained. Such an overall fragment pattern might be produced by either two different gene cluster copies, one having four (accommodating a snoR4 or any other snoRNA gene) and the other three genes, or by one cluster containing a second U31 and U51 gene, respectively. A second U51 gene can be ruled out because the U33/U51 primer pair produces only fragments of about 150 bp (Table 5.1). Although a second U31 gene and also any other snoRNA gene, cannot be excluded, it seems more likely that an *A. thaliana* gene order is maintained and that larger fragments (i.e. fragments of ca. 480 (U31/U51) and 340 bp (U31/U33), respectively; in bold, Table 5.1) host the snoR4 gene with primer site sequences that do not match the snoR4 primer designed (see also Chapter 4). While a U31/snoR4 gene order can be found in some other species (see Appendix, Figure A.16), a U31 tandem repeat is not known. However, it might be the case that these larger fragments (in bold, Table 5.1) do not contain any gene, but rather a long U31/U33 intergenic region. PCR amplification using a U31F/U31R primer pair might confirm or reject a U31 tandem repeat and sequencing would definitely settle the issue.

The estimated intergenic region for the ca. 290 bp U31/U51 fragment is zero (in italics, Table 5.1) suggesting that this fragment might not contain the U33 gene or only contains a fragment of the U33 gene due to partial gene loss. This is supported by the U31/U33 genotype profile, which does not contain an expected fragment of about 160 bp. Another U31/U51 fragment was approximately 190 bp long and, thus, can only consist of the genes U31 and U51.

Table 5.1: Estimation of intergenic regions and reconstruction of cluster A. In bold: intergenic region might contain a gene (e.g. snoR4). In italics: fragments which might not contain the U33 gene. The lengths of the amplified gene regions are shown in the table and the fragment sizes obtained for *Senecio* were rounded.

cluster A							
	Fragment sizes (bp)						
primer pair	*Arabidopsis*	*Senecio*	U31	intergenic	U33	intergenic	U51
U31/U51		*190*	*70*	*30*	*0*	*0*	*90*
		290	*70*	*0*	*90*	*40*	*90*
	445	340	70	50	90	40	90
	455	**480**	**70**	**190**	**90**	**40**	**90**
U31/U33*	324	200	70	50	80	-	-
	325	**340**	**70**	**190**	**80**	-	-
U33/U51	141, 150	150	-	-	20	40	90

* fragment sizes obtained by radio labelled genotyping only

From the fragments obtained in *Senecio* it appears that most snoRNA gene clusters are very similar in organisation to the gene clusters found in *A. thaliana*. However, gene losses, inversions, duplications, and differences in gene order relative to *Arabidopsis* were observed.

5.3.2 SnoRNA gene cluster organisation in *Senecio*

Cluster A

SnoRNA cluster A might be present in *Senecio* in four different copies with one copy containing the genes U31, most likely snoR4, U33 and U51, a second copy consisting of the U31, U33 and U51 genes, and a third and fourth copy containing only the U31 and U51 genes (Figure 5.1A).

Cluster B

Cluster B is most likely similar in structure to that found in *A. thaliana* and might also contain the box H/ACA snoR99 gene within the long intergenic region between the first two U14 genes (Figure 5.1B; Appendix, Figure A.17). Fragments with sizes of about 129/130, 680 and 694 bp were obtained, which represent only a subset of the fragments detected in *A. thaliana*. It is therefore feasible that the U14 genes within this cluster differ

slightly in primer sites and that primers only bind to the first, the third (U14-3) and the fourth (U14-4) U14 gene. Interestingly, while the length of the U14 genes was very constant among various species examined, the intergenic regions were highly variable and usually much longer than the calculated 40 bp for *Senecio* (see Appendix, Figure A.17). The similarity between *Senecio* and *Arabidopsis* of the short intergenic regions supports an *Arabidopsis* like U14 gene cluster organisation. The length difference between the longer fragments is most likely due to allelic variation rather than variation between different gene copies.

Cluster C

SnoRNA gene cluster C is present in at least two copies with intergenic regions of about 30 and 90 bp in length occurring between the U36 and U38 genes (Figure 5.1C and Appendix, Figure A.18). The U36 and U38 genes are of similar sizes across the majority of species.

Putative D cluster

Fragments that varied in length from 116 to 155 bp were obtained using the primer pairs U49/snoR2d and U49/snoR77Y. As these fragments are too short to contain the expected genes, this pattern is best explained by an inverted repeat of U49 (I-49; Figure 5.1D and Appendix, Figure A.19) which is present in some snoRNA cluster D copies. An extra inverted U49 gene was also identified in *Medicago truncatula* (see Appendix, Figure A.19). The reconstruction of cluster D using three primer combinations suggests one copy containing U49, an inverted U49, a long intergenic region which might accommodate snoR2, and snoR77Y (Figure 5.1D). Another copy might be similar to the gene cluster seen in *A. thaliana* consisting of the U49, snoR2 and snoR77Y genes. However, this gene cluster copy appears to be much shorter and it is therefore feasible that only the short U49 gene variant (~ 100 bp) is present in *Senecio*. At least one more copy consists of the U49 gene and its inverted repeat (Figure 5.1C).

Cluster E

Each of the primer pairs used for amplifying snoRNA cluster E appeared to amplify fragments from different regions of the cluster. It is deduced from this that the gene cluster is present in 3 copies with each copy having lost one of its constituent genes or, alternatively, the primer sites of constituent genes do not match primer sequences in all three copies.

Clusters F, G, I and M

The clusters F, G, I and M in *Senecio* show similar organisation to that identified in *A. thaliana*, but differ in copy number (Figure 5.1F, G, I and M and Appendix, Figure A.21Figure A.22Figure A.23). For example, while clusters F, G and M are single copy regions in *A. thaliana*, at least 2 copies of each of these clusters are found in *Senecio*. In contrast, whereas Cluster I appears to be present only once in *Senecio*, three copies of it occur in *A. thaliana*.

Cluster H

Cluster H was amplified using the internal U80 primers resulting in production of the expected fragment of 56 bp (Figure 5.1H). Surprisingly, a fragment of 138 bp and, in *S. aethnensis* only, one of 286 bp, were also generated suggesting a tandem repeat of one part and also the complete U80 gene (Figure 5.1H), respectively. A U80 tandem repeat was also detected in *Festuca pratensis* (see Appendix, Table A.28). The absence of the 286 bp fragment from any species other than *S. aethnensis* might be due to primer mismatch rather than the lack of this sequence.

Cluster J

Fragments generated by primer pairs of cluster J suggest an *A. thaliana* like gene cluster organisation in *Senecio* with one copy containing two and another copy containing just one snoR22 gene (Figure 5.1J and Appendix, Figure A.24). Evidence for the presence of the snoR80 gene in this cluster was obtained for a few samples using agarose gel electrophoresis that resolved fragments of the expected size.

Cluster N

While the snoR114/snoR85 primer pair produced only a few fragments of which two were similar to fragments detected in *Arabidopsis*, a more complicated fragment pattern was obtained with the snoR115/snoR85 primer combination. The large number and size of fragments obtained suggest complex gene organisations with duplications of individual genes. Although the fragment pattern of the snoR114/snoR85 suggests one *Arabidopsis* like gene cluster copy with an extra snoR114 and snoR85 gene (Figure 5.1N), the fragment pattern of the snoR115/snoR85 primer pair does not support the presence of a snoR115 gene within this cluster. However, the intergenic region calculated is large enough to harbour a snoR115 gene (Figure 5.1N). The other two gene cluster copies were reconstructed from snoR85 fragments. Both clusters contain three snoR85 genes and provide a long enough intergenic region between snoR114/snoR115 and the first snoR85 gene to accommodate an extra snoR114 and snoR85 gene (Figure 5.1N), respectively, but no fragment supporting their presence was obtained. The lack of fragments supporting the presence of the suggested genes in all copies might be due to primer mismatch. While clusters with three copies of snoR85, a snoR114-snoR85 gene order, and a cis-duplication of snoR114 were found in other species, no inversion of the snoR114-snoR115 gene order was evident (see Chapter 3, Figure 3.19).

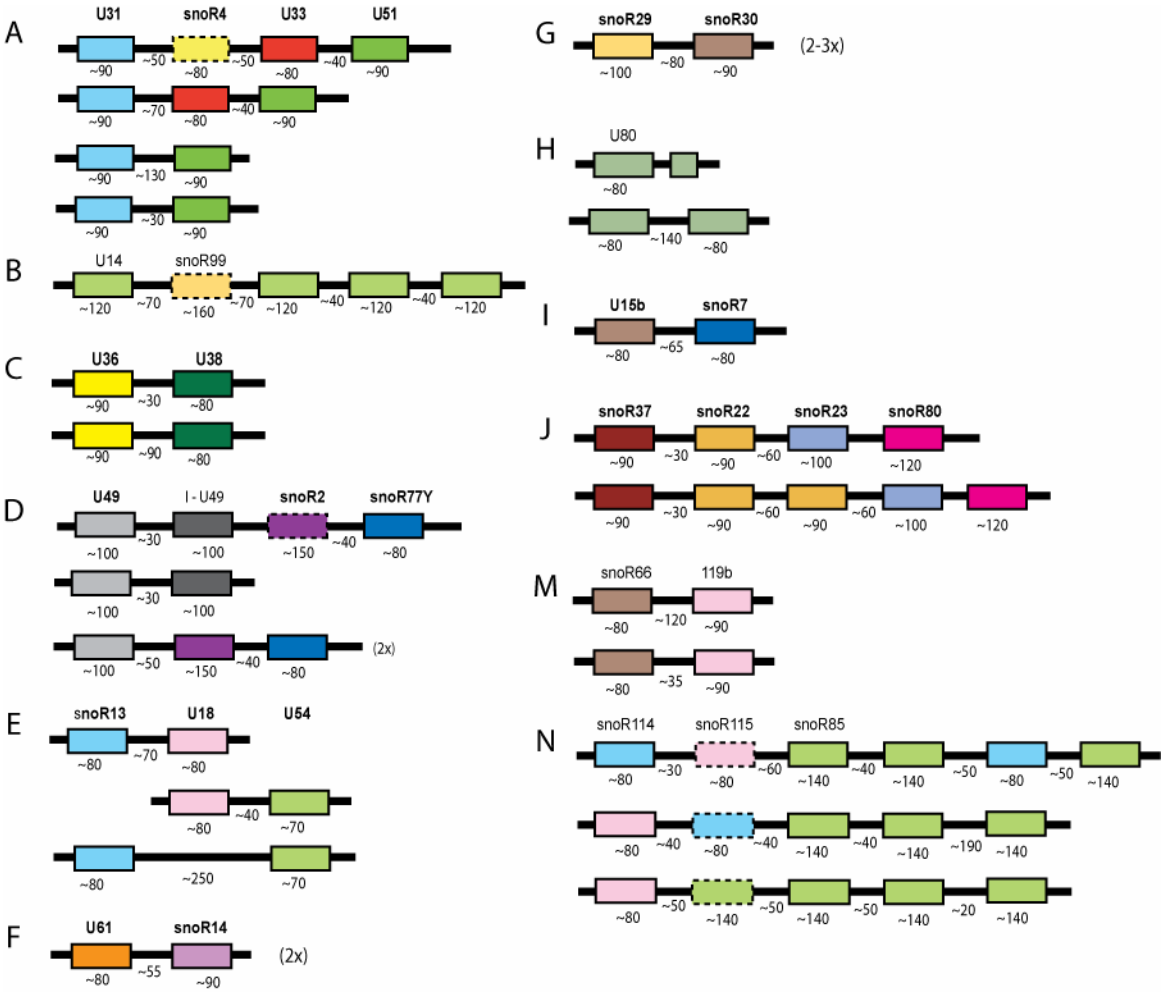


Figure 5.1: Proposed snoRNA gene clusters in *Senecio*. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the approximate lengths of genes and intergenic regions are indicated by numbers below boxes and lines, respectively. Dotted boxes represent putative genes which lack supporting fragment patterns (see text).

5.4 Discussion

Comparative analysis of snoRNA fragments obtained from *Senecio* species, using various snoRNA primer pairs, with putative fragments obtained by ePCR for *A. thaliana* and other species suggests that most of the amplified products represent snoRNA genes and snoRNA gene clusters. Furthermore, most clusters detected in *Senecio* appear to be organised similarly to related clusters in *A. thaliana*. However, some reconstructed clusters revealed differences in copy number and organisation.

5.4.1 Duplication and loss of snoRNA genes and gene clusters

In plants, about 50% of snoRNA genes are present in more than one copy. These different copies have arisen through duplication of complete gene clusters or parts of gene clusters (Brown *et al.*, 2003a). Polyploidisation is a major force in plant evolution and much of this duplication might be the result of hybridisation and genome duplication events in the evolutionary history of plants (Wendel, 2000). In this study, some snoRNA gene clusters were shown to be present in more copies in *Senecio* than in *A. thaliana*. For example, the snoR29/snoR30 gene combination of cluster G and the U61/SR14 gene pair appear to be present in at least two copies in *Senecio,* but in only one copy in *A. thaliana*. These gene duplications most likely happened after the split of the *Arabidopsis* and *Senecio* lineages and might suggest some hybridisation and/or genome duplication events after the origin of these gene clusters in the lineage leading to *Senecio*. However, one extra copy of each of snoR29 and snoR14 is also evident in *Arabidopsis* and, therefore, the differences in copy number might also be explained by the loss of the snoR30 and U61 gene within these copies. Both duplications and losses of genes and gene clusters play an important role in the evolution of snoRNA genes and gene clusters. While duplications give rise to new gene copies they are also responsible for gene losses due to gene redundancy, and as a consequence tolerance of mutations. However, these mutations can also give rise to novel snoRNA genes and are, therefore, important in the evolution of these genes (Brown *et al.*, 2001; Brown *et al.*, 2003a). Losses of gene clusters appear also to have happened

in *Senecio* relative to *A. thaliana*. For example, in *Senecio,* cluster I seems to be present only once, whereas two copies are found in *Arabidopsis*.

The number of genes and gene clusters was estimated from the number of different fragments based on length differences. Thus, it might be possible that more gene cluster copies with the same lengths are present and the number of gene cluster copies was underestimated in this study. Mismatches in the primer sequences of some genes would be another reason for differences in the number of gene/gene cluster copies detected between species although we would expect this effect to be minimal especially for box C/D snoRNA genes as primers are designed to sequences with complementarity to rRNA. An interesting example of gene cluster duplication, putative gene loss and gene evolution concerns cluster E. Cluster E appears to be present in three copies in *Senecio* with each copy seemingly to have lost a certain gene or at least its primer site. Most primers were designed using the rRNA antisense element and mutation in these sequences might indicate a novel methylation site and, thus, might give insight into the evolution of new genes.

5.4.2 Tandem gene duplication, inversions and inverted gene order

Most snoRNA genes in plants are organised in polycistronic clusters allowing coordinated expression. These clusters have arisen through tandem gene duplication followed by subsequent evolutionary change of these genes (Qu *et al.,* 2001; Brown *et al.*, 2003a). Therefore, tandem repeats play a major role in the evolution of gene clusters and their organisation. Cluster N reconstructed for *Senecio* differs in both the number of gene cluster copies and the number of snoR85 genes, most likely due to tandem duplication of genes detected in *Arabidopsis*. Furthermore, one copy might show an inverted gene order of snoR114 and snoR115 which could be explained by a duplication of the SnoR114 gene that is placed downstream of snoR115, followed by loss of the original snoR114 gene. Cluster D is represented by two copies which contain an inverted U49 (I-U49) gene. This is most like the result of a tandem repeat combined with an inversion of U49.

The reconstruction of gene clusters in *Senecio* reported in this chapter has been based on comparisons of fragment length pattern between various species. Further investigations using sequence information are necessary to confirm the patterns of gene cluster organisation that are proposed here, and to provide, in turn, a deeper insight into the evolution of snoRNA genes and gene clusters in *Senecio*.

Chapter 6: Sequence analysis of snoRNA genes and gene clusters in *Senecio squalidus* and related species

6.1 Introduction

Markers for a DNA barcoding system in plants should be short sequences that are easily amplified in many different species using universal primers. In addition, sequence variation for such markers should be much greater between species than within species and represent single or low copy genes. As demonstrated in the previous chapters of this thesis, the majority of universal snoRNA primers used on *Senecio* amplify short putative snoRNA genes/gene clusters in the genus. Sequencing of fragments generated by these primers should (i) confirm that the correct snoRNA genes/gene clusters were amplified, (ii) show the level of sequence variation present within and between species, and (iii) indicate how many different gene copies (paralogues) of a certain gene/gene cluster might be present in *Senecio*. Furthermore, sequencing should reveal differences in snoRNA gene organisation in *Senecio* species relative to *Arabidopsis thaliana*. In the research reported in this chapter, a subset of possible snoRNA gene clusters, previously detected using fragment analysis (Chapter 5), was subjected to analysis of sequence variation (Figure 6.1). According to the *A. thaliana* genome this subset should include one gene cluster that is present in two copies (Cluster A; Figure 6.1A), a gene cluster comprising four homologous box C/D genes and a very recently identified box H/ACA gene (Cluster B; Figure 1B), plus two single copy gene clusters (Clusters F and G; Figure 6.1F and G). Fragment analysis indicated that within *Senecio* there are more copies of these genes/gene clusters relative to the number of copies present in *A. thaliana.*

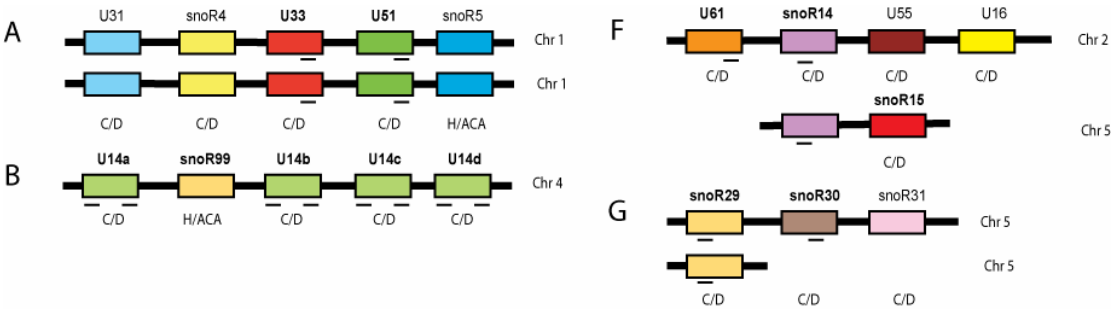


Figure 6.1: Gene organisation of snoRNA genes and gene clusters in *Arabidopsis thaliana* that were subjected to sequence analysis in *Senecio*. Gene clusters are indicated by capital letters. The approximate locations of universal primer sites used for sequencing are indicated by black lines below genes. Genes are displayed by boxes of different colours with their names written above and their chromosome number to the right. The labels C/D and H/ACA below genes indicate the snoRNA gene type.

6.2 Material and Methods

6.2.1 Plant Material

Leaf material for DNA extraction was obtained from plants cultivated from seed in the greenhouse. These included samples of species previously used for analysis (Chapters 4 and 5) plus 41 further samples comprising *S. aethnensis* (10 samples), *S. chrysanthemifolius* (one sample), natural *S. aethnensis* x *S. chrysanthemifolius* hybrids (20 samples), artificially produced *S. aethnensis* x *S. chrysanthemifolius* F_1 hybrids (5 samples) and the reciprocal *S. chrysanthemifolius* x *S. aethnensis* F_1 hybrids (5 samples) (see Chapter 2, Table 2.1).

6.2.2 DNA-Extraction and PCR-amplification

Total DNA was extracted from either frozen or fresh leaves. Leaf tissue was pulverized to a fine powder using liquid nitrogen and DNA was isolated using a modified 2 x CTAB (hexadecyltrimethyl ammonium bromide) extraction method (Doyle & Doyle, 1987). Samples were amplified by PCR (see Chapter 2) using four universal primer combinations (snoR29/snoR30, U14-1/U14-2, U33/U51 and U61/snoR14, see Table 2.2),

and also 24 newly designed specific primers (Table 6.2, Table 6.3 and Table 6.5; see also Table 2.2).

6.2.3 Sequencing

For the primer pairs U14-1/U14-2 and U33/U51, PCR-products of a single individual were purified, cloned and sequenced as described in Chapter 2. For the other universal primer combinations, snoR29/snoR30 and U61/SR14, and also for some specific snoR29/snoR30 primer pairs, samples of the same species were pooled prior to purification. Alternatively, specific snoR29/snoR30 PCR-products of single samples were sequenced directly after PCR without cloning.

6.2.4 Molecular data analysis

Target sequences were aligned using CLUSTAL W algorithm (Thompson *et al.*, 1994) incorporated either in SEQUENCHER (Gene Codes Corporation) or BioEdit 7.0.9.0 (Hall, 1999) and alignments were visually improved. BLAST searches (see Chapter 3) were performed and sequences of *Senecio* species, if found, were included in analyses. The basic molecular characteristics of different sequences were examined using the programme MEGA 4.1 (Tamura *et al.*, 2007).

6.2.4.1 Indels and Missing data

Indels (insertions/deletions) can be a valuable source of phylogenetic information. Despite their potential benefits their application might be difficult because they might be unreliable as characters due to difficulties in determining character states based on indels and in defining the homologous states of gaps (Young & Healy, 2003; Pons & Vogler, 2006; Simmons *et al.*, 2007). Thus, although phylogenetic information might be lost, indels in alignments of cloned sequences were treated as missing data and subsequently supported by pairwise deletion (e.g. MEGA and PAST 1.99 (Hammer *et al.*, 2001)) or interpolating sample-by-sample pairwise distances (e.g. GenAlEx 6.3 (Peakall & Smouse, 2006)). However, sequences produced by direct sequencing were shorter (one primer and

adjacent nucleotides not readable) and therefore missing data at one end emerged when these sequences were included in alignments. These missing data were excluded from the dataset analysed (complete deletion option).

6.2.4.2 Genetic distance analysis

Principal coordinate analyses (PCO) were conducted on datasets using Euclidian (in PAST) and Nei's genetic distances (in GenAlex 6.3). Neighbour Joining (NJ) analyses (Saitou & Nei, 1987) were performed using the maximum likelihood composition model and Euclidean genetic distances as incorporated in the programmes MEGA 4.1 and PAST 1.99, respectively. To estimate support of tree nodes, bootstrap values (Felsenstein, 1985) were calculated from 1000 pseudoreplicates.

6.2.4.3 Identification of putative gene copies

Putative gene copies are assumed to be more divergent in their sequences than alleles and, as already mentioned in Chapter 4, they might also differ in length of sequence. Thus, different gene copies might form well-separated groups/well-supported clades which might also show differences in sequence length. To identify putative gene copies, the lengths of sequences that formed a well-supported clade in a NJ tree were examined and compared with the results obtained from fragment analysis (Chapter 4). Sequences of certain length found in one particular clade might be used to assign fragments found in the fragment profiles to this group. Thus, an NJ tree might consist of two well-supported clades, one containing sequences of size A the other sequences of size B. If these clades represent different gene copies then each fragment profile would show both sizes. In diploid species the number of fragments per sample assigned to a certain clade cannot be more than two for a single copy of a gene (i.e. maximum of two alleles per gene). Thus, a higher number of fragments per sample per clade would indicate that sequences were amplified from different gene/gene cluster copies. This approach was used to assign the fragments of the fragment analyses matrices of snoR29/snoR30 and U61/snoR14 to different clades obtained by sequence analysis.

6.2.4.4 Specific primer design

Universal primer pairs might cause the amplification of possible paralogous regions and, therefore, primer modification might be neccessary to isolate and investigate these paralogues in more details. Additionally, these modified (specific) primers might amplify single copy regions which, although not applicable for DNA-barcoding, might be useful markers for phylogenetic investigations. Thus, primers were designed to amplify different clusters and putative gene copies, respectively. The sequences generated were aligned and identical sequences were removed and sorted according to groupings revealed by PCO and NJ analysis. Afterwards, alignments were examined for polymorphic and group specific sites (i.e. polymorphisms that distinguish groups) and primers were designed accordingly. Some primers were examined for amplification success only (e.g. U14) while others were used for fragment analysis (e.g. snoR29/snoR30) and/or sequencing (e.g. U61/snoR14, snoR29/snoR30). New sequence data were added to the existing alignments that were then re-analyzed.

6.3 Results

PCR amplification products of four different universal snoRNA primer pairs were subjected to sequence analysis to confirm their snoRNA gene/gene cluster origin. All of the 'original' primer pairs generated some degree of sequence variation and more than one gene copy across samples (Table 6.1). Specific primers were designed for all sequences generated, except for those generated by U33F/U51R, so as to allow further investigation of sequence variation.

Table 6.1: Variation and gene copies of the snoRNA gene clusters investigated.

Cluster	Primer pair	Variation	Multiple copies	Specific primers
A	U33F/U51R	+	+	-
B	U14-1/U14-2	+	+	U14-2 variants
F	U61F/snoR14R	+	+	U61F variants
G	snoR29F/snoR30R	+	+	snoR30R variants

Nucleotides that distinguished putative gene copies were positioned adjacent to at least one of the original universal primer sequences used in the analysis and, therefore, all specific primers were designed by elongating an original universal primer sequence. In some cases the complete sequence of an original primer was used in this process, whereas in other cases only a part of it was incorporated in the specific primer generated. Furthermore, it was only necessary to use one specific primer type, forward or reverse, in further sequence analysis (see below).

6.3.1 U33/U51

6.3.1.1 Sequence generation from original universal primers

From two samples of each of *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus* and *S. cambrensis*, and one sample of *S. vulgaris*, 112 clones yielded 102 good quality sequences of about 150 bp in length. These comprised 11 sequences of *S. aethnensis*, 32 of *S. chrysanthemifolius*, 29 of *S. squalidus*, 8 of *S. vulgaris*, and 22 of *S. cambrensis*. The number of different sequences identified in a particular sample ranged from two (in a sample of *S. aethnensis*) to eight (in a *S. cambrensis* sample). However, the maximum number of different sequences obtained for a diploid species (in samples of *S. chrysanthemifolius*) was six, indicating that at least three copies of this gene combination are present in diploid *Senecio*.

After removing all identical sequences, the alignment consisted of 21 different sequences having a maximum length of 153 bp. Of this, 57 bp comprised the intergenic region, and 96 bp comprised the gene region which included both primer sequences (Figure 6.2). The intergenic region, which is underlined in red in Figure 6.2, contained 19 variable sites and 5 indels, whereas the U51 gene region (3' end gene; underlined in black in Figure 6.2) contained 8 variable sites and 1 indel. The U33 gene region (5' end gene, underlined in black in Figure 6.2) contained only its primer sequence and therefore no variable sites were observed (Figure 6.2). It should be noted that both *S. squalidus* samples and one individual of *S. cambrensis* examined contained sequences that differed in their box C sequences by one nucleotide from all other sequences.

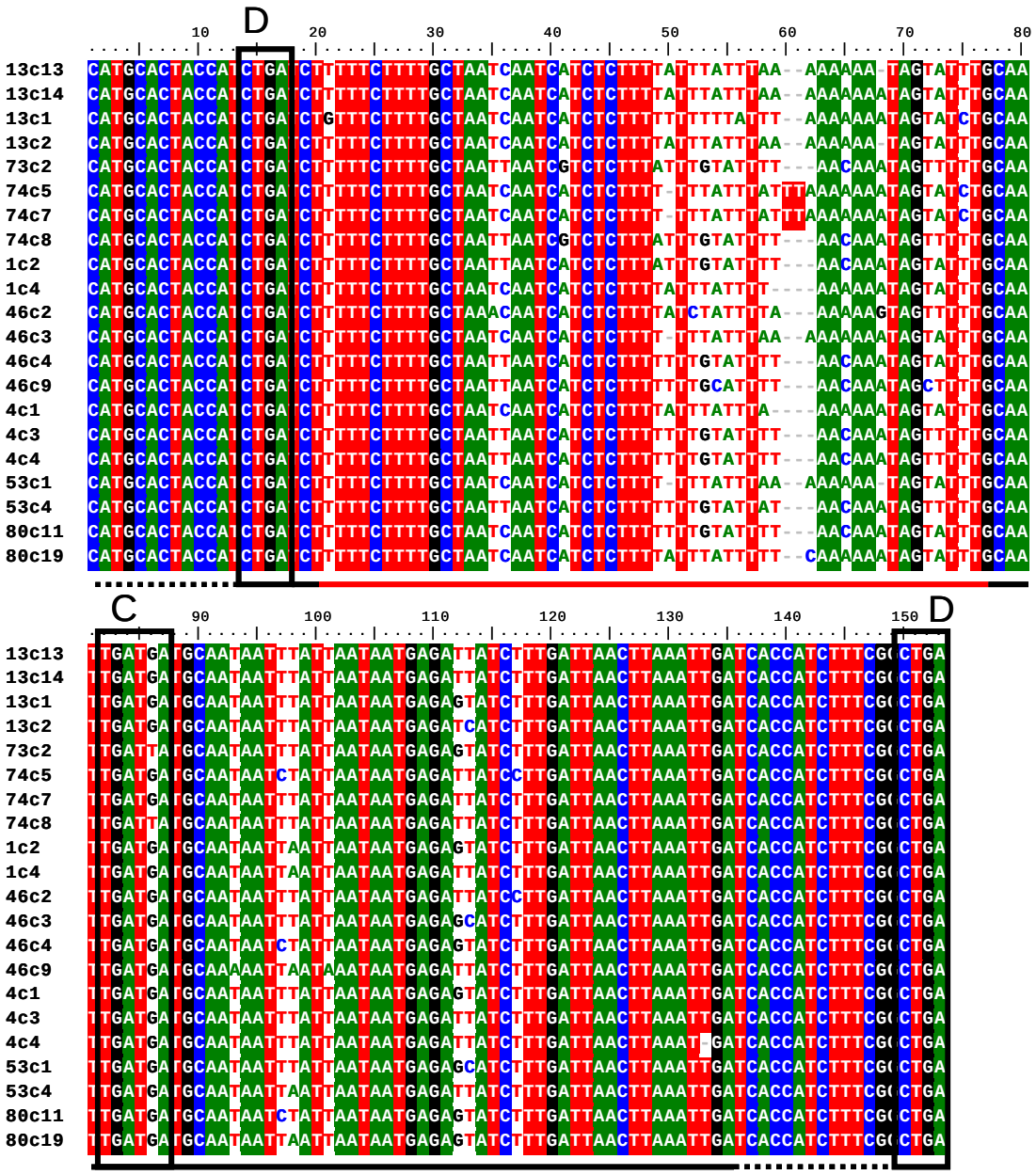


Figure 6.2: Alignment of 21 different U33/U51 sequences in *Senecio*. Conserved positions are shaded. Black line = gene region; black dotted line = antisense element; red line = intergenic region; C and D boxes are indicated (C and D). Sequence names refer to the sample and clone they were taken from. 1, 4 = *S. aethnensis*; 13 = *S. chrysanthemifolius*; 73, 74 = *S. squalidus*; 46 = *S. vulgaris*; 53, 80 = *S. cambrensis*. Some identical sequences were found in more than one sample.

6.3.1.2 Gene copies and organisation

The different U33/U51 sequences identified fell into two clades (indicated by green and red bars) in the NJ tree (Figure 6.3). Both clades were structured containing several highly supported subclades (Figure 6.3). It was noted that almost all sequences of *S. vulgaris* differed considerably, occupying single clades or sharing with *S. cambrensis* only. *Senecio cambrensis* is the allohexaploid hybrid of *S. vulgaris* and *S. squalidus*, and possessed all of the U33/U51 sequences found in *S. vulgaris* and some of those present in *S. squalidus*. It was also the case that the homoploid hybrid species *S. squalidus* contained all of the sequences found in one of its parents, *S. chrysanthemifolius*, and some of those present in its other parent, *S. aethnensis* (Figure 6.3).

All species except *S. vulgaris* were represented by two samples in this study. Rather surprisingly, sequences found within individuals were placed in different clades indicating that these sequences might represent different gene copies. However, more than two sequences obtained from samples of diploid species (e.g. *S. chrysanthemifolius* (sample 13) and *S. squalidus* (sample 74)) were placed within the "green" clade suggesting more than one gene copy present in this clade. Interestingly, some sequences obtained from samples of the same species differ greatly whereas others show complete identity. For example, while the sequences of the two *S. aethnensis* samples differed considerably, *S. cambrensis* samples share four sequences (Figure 6.3).

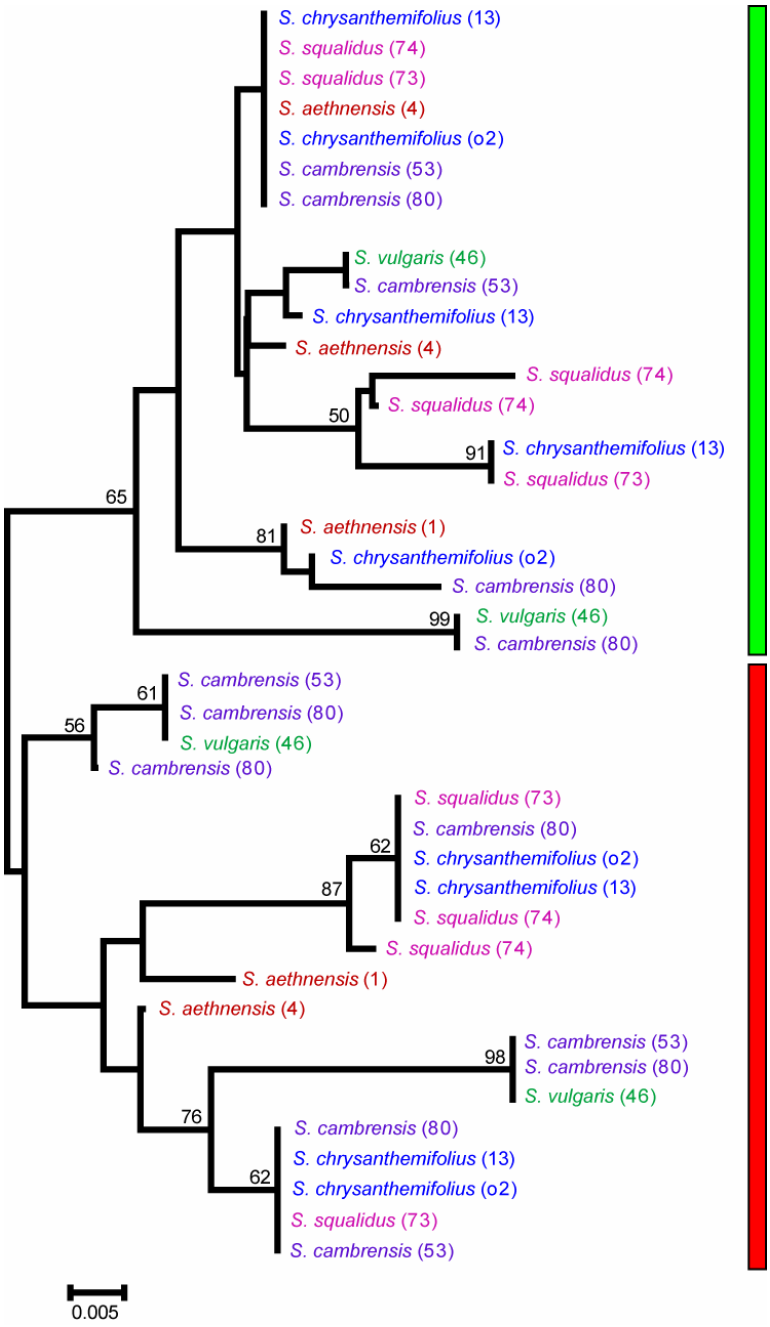


Figure 6.3: NJ tree derived from different sequences generated by the U33F/U51R primer pair across five species of *Senecio*. Relationships are based on sequence variation using the Maximum Composite Likelihood method and pairwise deletion option for gaps/missing data. Bootstrap values are shown above or below branches. Sample identification numbers are given in brackets after species' names. The coloured vertical bars indicate the different clades and therefore the putative gene copies present across the *Senecio* species surveyed.

6.3.2 U14-1/U14-2

6.3.2.1 Sequence generation from original primers

The U14-1/U14-2 gene region was examined by sequencing 154 clones from individual samples (CT, Figure 6.4), including *S. aethnensis*, *S. vulgaris*, *S. chrysanthemifolius*, *S. squalidus* and *S. cambrensis* (Figure 6.4). From these 154 clones, 136 good quality sequences were obtained (RS, Figure 6.4), which differed in length and contained either one, two, three or four U14 genes (1G, 2G, 3G and 4G, Figure 6.4). Most clones (94 sequences), contained just one gene sequence and clones comprising three genes were rare (7 sequences). Only one clone (from *S. aethnensis* (sample 8)) was found to contain four U14 sequences. Interestingly, longer sequences containing more than one gene were obtained more often from diploid species than from the tetraploid/hexaploid species. For example, in *S. vulgaris* only two of the 43 clones examined contained two genes, while in *S. cambrensis* all clones consisted of only one gene. In contrast, in both *S. chrysanthemifolius* samples examined the majority of clones contained two rather than one gene sequence. However, it should be noted that this apparent distribution of fragments containing one, two, three and four genes, respectively, was most likely caused by preferential PCR amplification and cloning rather than this having any biological or evolutionary significance.

After the extraction of all U14-1/U14-2 sequences from longer sequences (i.e. clones containing more than one U14-1/U14-2 gene sequence) a total number of 187 U14 gene sequences were aligned (SGS, Figure 6.4). The number of different sequences present within diploid samples differed greatly and ranged from seven in *S. aethnensis* (2) to 17 in *S. chrysanthemifolius* (o2)) (DS, Figure 6.4). Thus, in *Senecio*, at least 9 different gene copies (17 different sequences in *S. chrysanthemifolius* (sample o2)) are organized in gene clusters containing up to at least four genes.

sample	CT	RS	1G	2G	3G	4G	SGS	DS
S. aethnenis (2)	8	7	5	2	0	0	9	7
S. aethnenis (4)	8	8	7	1	0	0	9	7
S. aethnenis (8)	16	13	5	4	3	1	26	12
S. chrysanthemifolius (o2)	16	14	5	8	1	0	24	17
S. chrysanthemifolius (13)	16	12	4	7	1	0	21	14
S. squalidus (74)	16	13	8	5	0	0	18	12
S. squalidus (76)	14	13	7	5	1	0	20	14
S. vulgaris (46)	8	8	6	2	0	0	10	7
S. vulgaris (47)	16	14	14	0	0	0	14	14
S. vulgaris (81)	8	8	7	0	1	0	10	8
S. cambrensis (53)	8	7	7	0	0	0	7	3
S. cambrensis (55)	8	8	8	0	0	0	8	6
S. cambrensis (80)	12	11	11	0	0	0	11	7

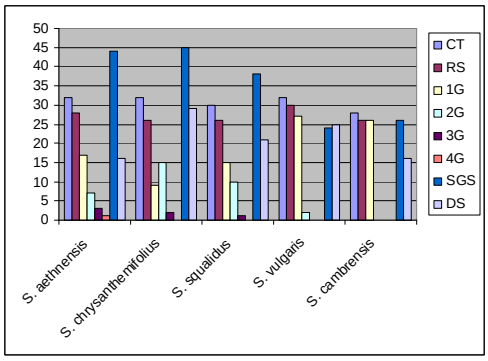


Figure 6.4: U14-1/U14-2 sequences obtained from 154 clones. Thirteen samples from five species were analysed yielding 136 readable sequences. These sequences contained either 1, 2, 3 or 4 U14 genes. CT = clones tested, RS = readable sequences, 1G to 4G = one to four genes, SGS = Sum of gene sequences, DS = different sequences. Sample numbers are in brackets.

After removing identical sequences from the alignment the dataset consisted of 72 different sequences of 95 bp with 31 variable sites and 10 indels (Figure **6.5**). It should be noted that the U14-1/U14-2 sequence represents approximately only the final 85 bp part of the gene which is normally about 125 bp long.

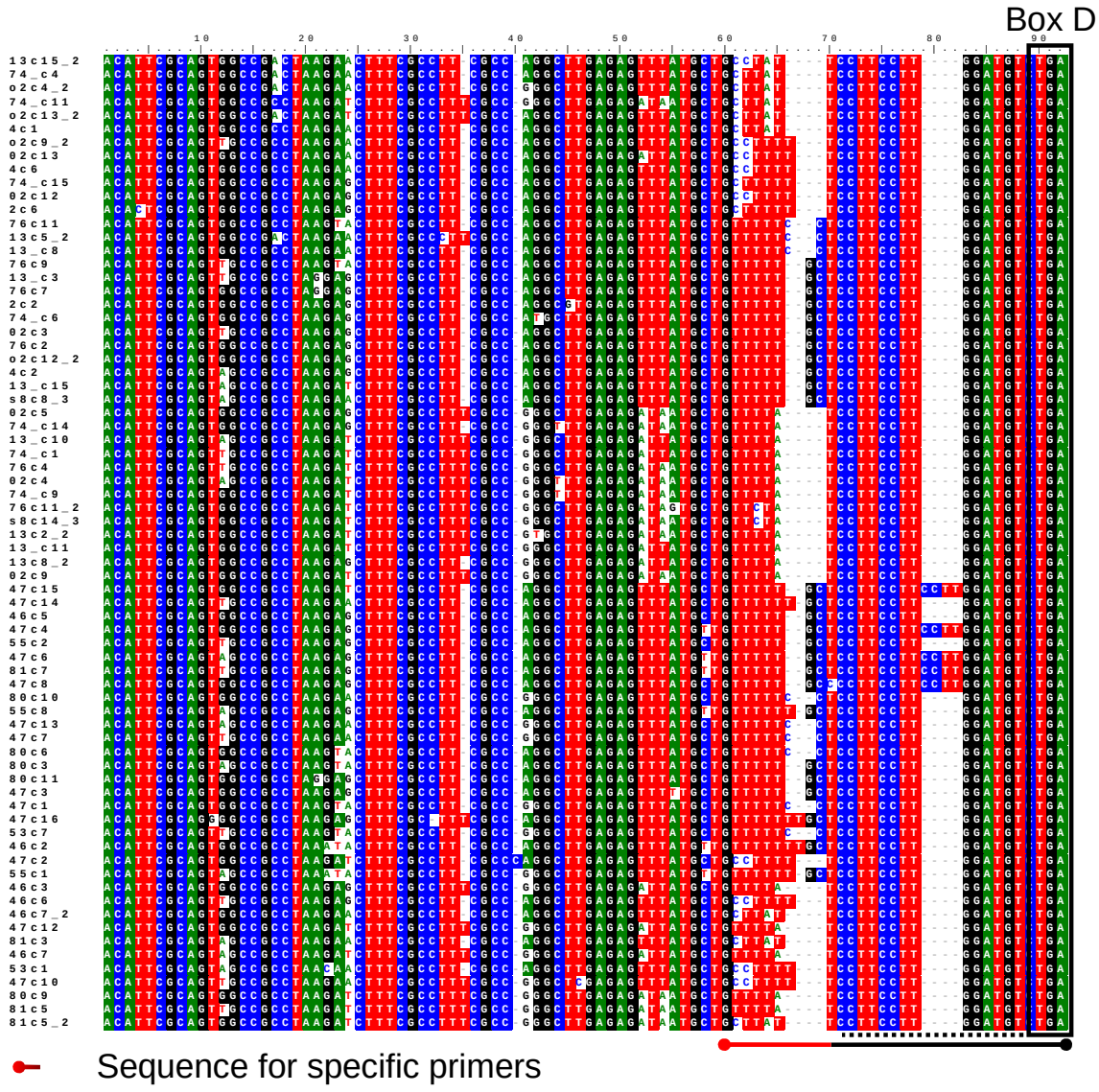


Figure 6.5: Alignment of 72 different U14-1/U14-2 sequences from thirteen samples of five *Senecio* species. Conserved positions are shaded. The red part of the specific primer site symbol indicates the extension of the original primer. Dotted line = antisense element; box D is indicated. Sequence names refer to the sample and clone they were taken from. 2, 4, 8 = *S. aehtnensis*; 13, o2 = *S. chrysanthemifolius*; 74, 76 = *S. squalidus*; 46, 47, 81 = *S. vulgaris*; 53, 55, 80 = *S. cambrensis*. (see also Figure 6.4).

6.3.2.2 U14 gene copies and organisation

A total of 188 sequences (including one *S. chrysanthemifolius* sequence found by BLAST search - gi|89504665), were placed into one of four groups distinguished by PCO analysis (red, green, blue and violet encircled, Figure 6.6). However, none of these groups received high bootstrap support in a NJ tree (not shown). Each group, except one, contained sequences from each of the five species (Figure 6.6) and also each sample (not shown). Within each group the species were intermixed.

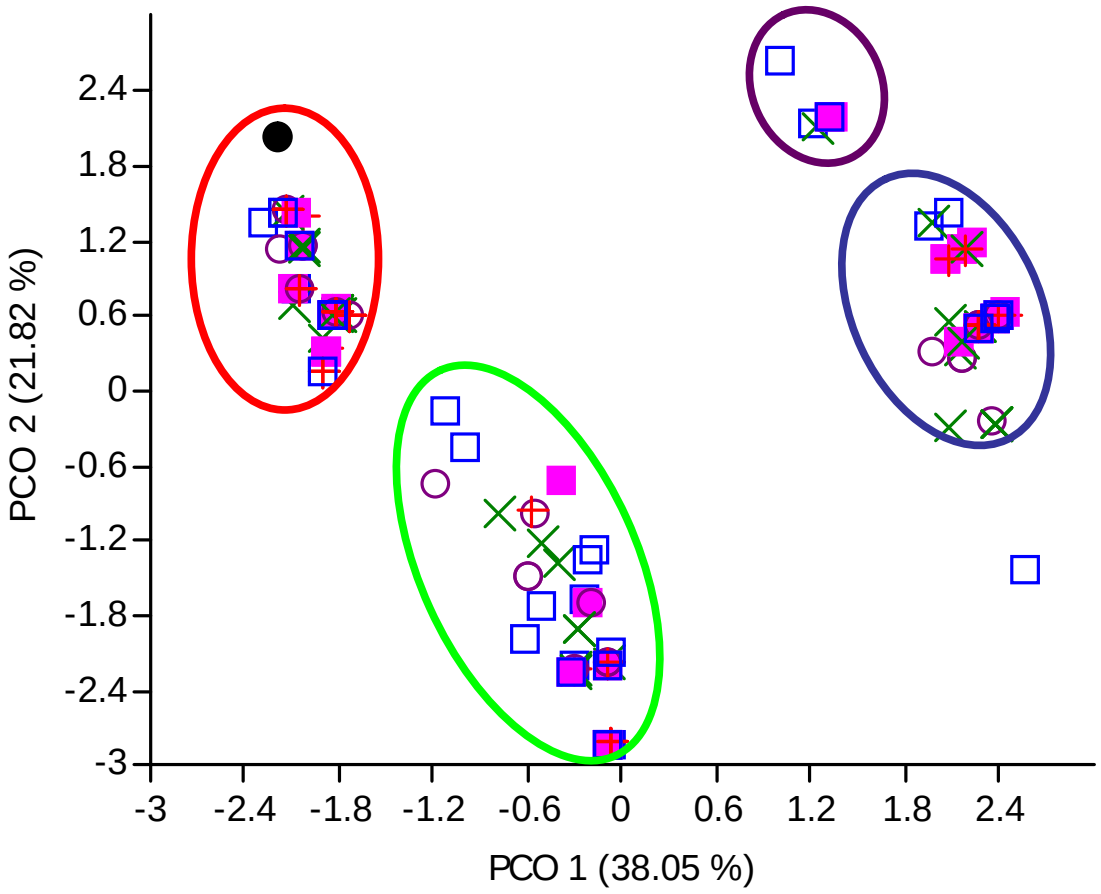


Figure 6.6: PCO plot based on an analysis of 188 U14 gene sequences. *S. aethnensis* = +; *S. chrysanthemifolius* = □; *S. squalidus* = ■; *S. vulgaris* = **x**; *S. cambrensis* = ○; gi|89504665 = ●.

The different U14 gene copies extracted from longer sequences (i.e. clones containing 2, 3 and 4 gene copies, respectively) were numbered according to their 5'-3' order and their positions in a NJ tree were compared. The grouping patterns (i.e. positions of various gene copies extracted from one clone within a NJ tree) differed for most of the longer sequences obtained. An example of a NJ tree constructed from all *S. aethnensis* (sample 8) gene sequences is shown in Figure 6.7. The gene copies of various clones (c1 to c16) are indicated by the numbers given (e.g. 8c6 and 8c6_2: first and second gene sequence of clone 6 from the *S. aethnensis* sample 8). While some of these copies differ considerably and were placed in different positions within the tree, other sequences were highly similar. For example, the clone 8c8 (light blue, Figure 6.7) contained sequences of three copies of the U14 gene, the first one (8c8) is found at the top, the second one (8c8_2) at the base and the third copy (8c8_3) in the middle of the tree. Three identical sequences were obtained from the 8c13 clone (dark green, Figure 6.7) and were positioned in the middle of the tree (8c13, 8c13_2 and 8c13_3). Some gene sequences of different clones were identical but showed differences in their other gene copies which might be placed at different positions within the tree. For example, the clone 8c16 (purple, Figure 6.7) contains three genes, the first (8c16) and the third (8c16_3) are identical and found at the tree basis and the second (8c16_2) in the middle of the tree. The second gene of 8c8 is identical to the former two but its first and third gene can be found in different positions than the second 8c16 gene. Overall, the differences in the gene copy grouping patterns of various clones indicate a much higher number of U14 gene copies present in *Senecio* than suggested by U14-1/U14-2 sequence variation.

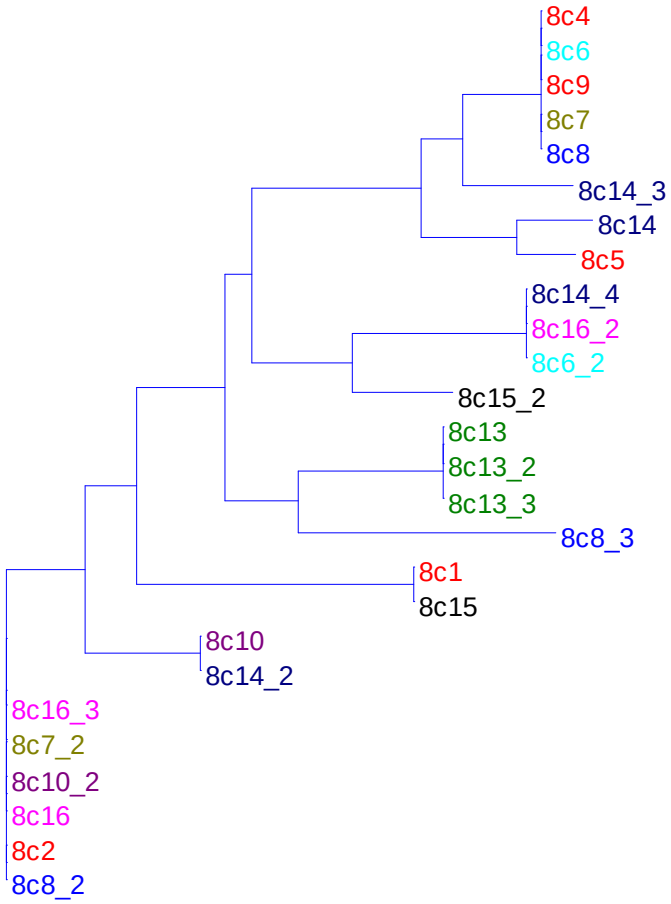


Figure 6.7: NJ tree of all U14 gene sequences extracted from one sample (*S. aethnensis* (sample 8)). 8 = sample 8. c1 to c16 = clone 1 to 16. Clones which consisted of one U14-1/U14-2 sequence in red. Different colours refer to different clones which contain at least two gene copies indicated by numbers (e.g. 8c13 = gene 1, 8c13_2 = gene 2 and 8c13_3 = gene 3).

With the exception of *S. cambrensis* (for which no sequence containing more than one U14 gene was obtained (see Figure 6.4)), U14 genes were usually separated in clusters by conserved intergenic regions of about 50 bp, although in some cases longer intergenic regions were evident ranging from 190 bp in *S. vulgaris* (46) to 229 bp in *S. chrysanthemifolius* (o2) and *S. aethnensis*. Intergenic regions of 206 and 221 bp in length were present in *S. squalidus* (74) and *S. chrysanthemifolius* (13) samples, respectively. An alignment of the seven long intergenic regions detected (Figure 6.8) showed highly conserved regions including an H and ACA box sequence which could be identified as

box H/ACA snoR99 gene (Figure 6.8) which is placed between the first two U14 genes, U14a and U14b, respectively, in *A. thaliana* (see Figure 6.1B). Although the three *S. squalidus* (74) intergenic sequences were identical, their adjacent 3' genes differed and, thus, the snoR99 gene is present at least twice in *Senecio*.

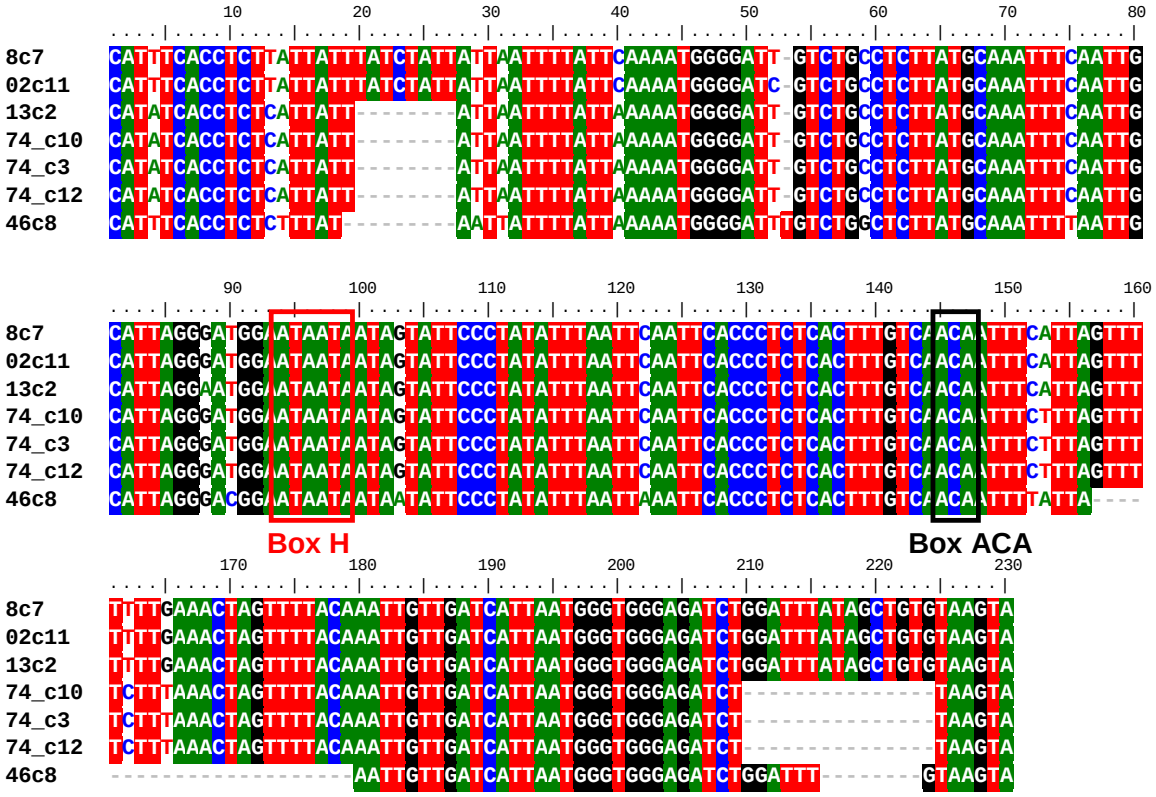


Figure 6.8: Alignment of 7 long intergenic U14 sequences. Conserved positions are shaded. These sequences were obtained from five different samples made up of four species and contain the box H/ACA snoR99 gene. H and ACA boxes are indicated. Sequence names refer to the sample and clone they were taken from. 8 = *S. aethnensis*; 13, o2 = *S. chrysanthemifolius*; 74 = *S. squalidus*; 46, = *S. vulgaris*.

6.3.2.3 Design of clade/gene specific primers

Fourteen different reverse U14-2 primers with lengths between 25 and 29 nt, TMs ranging from 54.4 to 58.7 °C, and GC contents from 40 to 48 % (Table 6.2) were designed by elongation of the universal primer (see alignment Figure 6.5). Five primer sequences (U14-2.1 to U14-2.5, Table 6.2) differed considerably from each other and

might be used to amplify putative gene copies placed in different groups in the PCO analysis (see Figure 6.6). The remaining primers (indicated by the letters a to c, Table 6.2) are variants of these five sequences and might be useful in distinguishing putative gene copies within the same group.

Table 6.2: **Sequences and characteristics of specific primers designed for amplifying different U14 snoRNA gene cluster sequences.** The first primer shown is the original universal primer sequence. a, b, c = variants of group specific primers.

Name	Direction	Sequence (5' - 3')	Length (nt)	TM (° C)	GC (%)
U14-2		TCAGACATCCAAGGAAGGA	19	48.9	47.7
U14-2.1		TCAGACATCCAAGGAAGGAATARGC	25	56-57.7	44
U14-2.1a		TCAGACATCCAAGGAAGGAATAAGC	25	56	44
U14-2.1b		TCAGACATCCAAGGAAGGAATAGGC	25	57.7	48
U14-2.2		TCAGACATCCAAGGAAGGAAAAARGC	26	56.4-58	42.3
U14-2.2a		TCAGACATCCAAGGAAGGAAAAAAGC	26	56.4	42.3
U14-2.2b		TCAGACATCCAAGGAAGGAAAAAGGC	26	58	46.2
U14-2.3	reverse	TCAGACATCCAAGGAAGGAGCAAAAA	26	56.4	42.3
U14-2.3a		TCAGACATCCAAGGAAGGAGCAAAAAC	27	58.2	44.4
U14-2.3b		TCAGACATCCAAGGAAGGAGCAAAAAAC	28	58.5	42.9
U14-2.3c		TCAGACATCCAAGGAAGGAGCAAAAAAAC	29	58.7	41.4
U14-2.4		TCAGACATCCAAGGAAGGATARAAC	25	54.4-56	40
U14-2.4a		TCAGACATCCAAGGAAGGATAAAAC	25	54.4	40
U14-2.4b		TCAGACATCCAAGGAAGGATAGAAC	25	56	44
U14-2.5		TCAGACATCCAAGGAAGGAGGAAAAAC	27	58.2	44.4

6.3.2.4 Amplification by specific snoRNA primers

All specific U14 primers (Table 6.2) together with the U14-1 primer were tested for PCR amplification using one sample of *S. aethnensis*, two of *S. chrysanthemifolius* and one of *S. squalidus*. The expected band of about 80 bp was obtained from all reactions (not shown). Furthermore, some primers, e.g. U14-2.2, U14-2.3 and U14-2.4, amplified additional bands of about 250, 450 and 750 bp in length, respectively. However, differences in banding patterns were evident between different variants (a, b and c; Table 6.2). For example, while primer U14-2.3a produced strong bands of 250 and 750 bp in both *S. chrysanthemifolius* samples, primer U14-2.3b amplified a different band of about 450 bp in length. Unsurprisingly, all three of these bands were generated by the U14-2.3 primer. There were also differences in amplified band patterns between samples and species with both *S. chrysanthemifolius* samples exhibiting stronger band amplification

than was the case in either of the other species. Additionally, bands present in both samples of *S. chrysanthemifolius* were sometimes absent from one or both of the other species. For example, only one band was generated by the U14-2.3 primers in *S. aethnensis* and *S. squalidus*.

In summary, the U14-1/U14-2 universal primer pair produced sequences of multiple gene copies (at least nine in one diploid), but it was not possible to distinguish between alleles, paralogues and orthologues of the U14 gene. Furthermore, the sequences obtained from different samples/species were intermixed into four different groups resolved by PCO analaysis. These four groups did not distinguish species from each other, but rather identified four different types of sequence. Thus, sequences of U14 genes placed within one particular group were either identical or very similar to each other, whereas those placed in different groups differed considerably from one other. Some intergenic regions were found to be relatively long and to accommodate the snoR99 box H/ACA gene. Specific primers were designed and amplified successfully some samples, but were not further investigated because of the high complexity of this gene cluster. Further examination of this cluster would be time consuming and was, therefore, not possible within the scope of this thesis.

6.3.3 U61/SnoR14

6.3.3.1 Sequence generation from original universal primers

Sequences generated by the U61/SR14 primer pair were obtained from three species (*S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus*) after pooling eight to ten samples per species. The initial data matrix comprised 89 sequences, i.e., 38 from S. *aethnensis*, 12 from *S. chrysanthemifolius* and 39 from *S. squalidus*, with each sequence being 137 bp in length. Identical sequences were removed to leave 46 sequences in the alignment (Figure 6.9). The intergenic region (71 bp) contained 36 variable sites, while the two gene regions (66 bp) had seven variable sites. Only one indel (1 bp) was present in the snoR14 gene region, while several indels up to a maximum length of 12 bp were evident in the

intergenic sequence. Interestingly, the indel of the snoR14 gene was present in its box C, which showed higher variability than other parts of the same gene sequence (Figure 6.9).

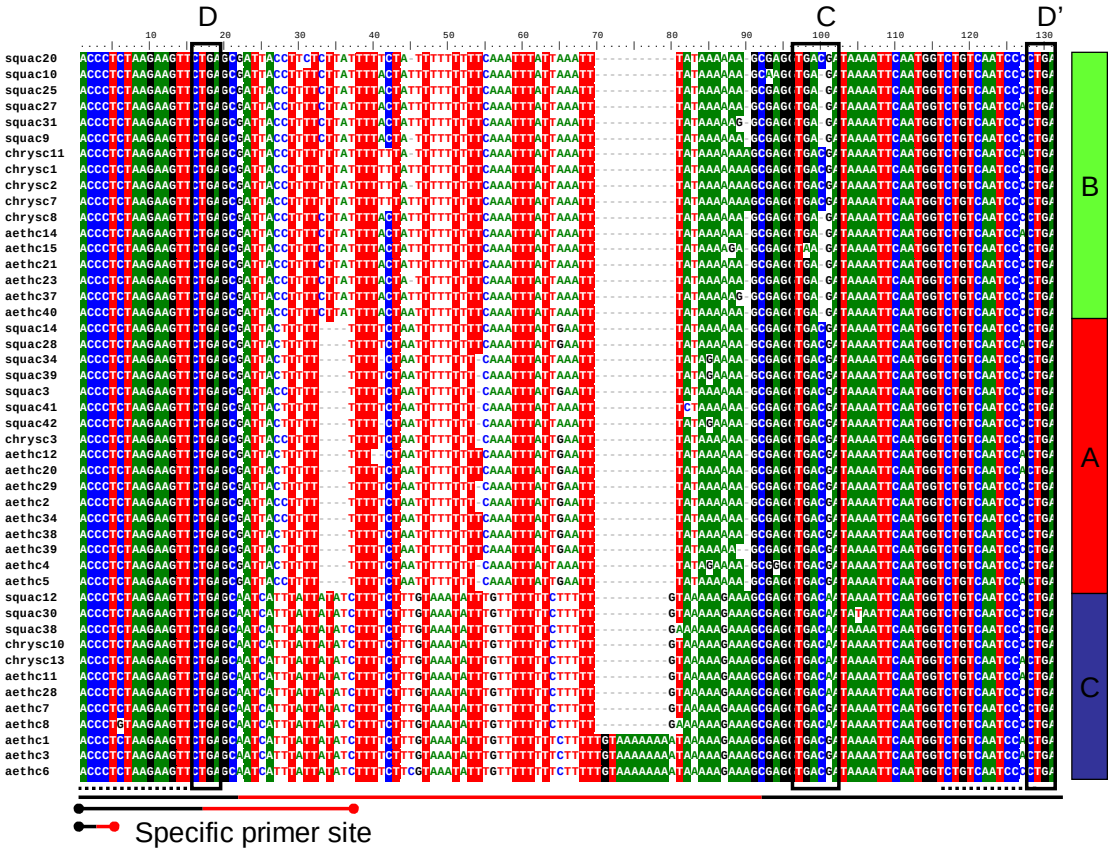


Figure 6.9: Alignment of 46 different U61/snoR14 sequences detected across three *Senecio* species. Conserved positions are shaded. The red part of the specific primer site symbol indicates the extension of the original primer. Coloured vertical bars indicate the clusters (A, B and C) to which the sequences belong to in the PCO plot and NJ tree (see below). Dotted line = antisense element; black line = gene region; red line = intergenic region. aeth = *S. aethnensis*, chry = *S. chrysanthemifolius*, squa = *S. squalidus*.

6.3.3.2 U61/snoR14 gene copies and organisation

Three well separated groups of U61/snoR14 sequences were detected by means of PCO analysis (Figure 6.10). Sequences obtained from each species (*S aethnensis*, *S. chrysanthemifolius* and *S. squalidus*) were present in each of these three different groups.

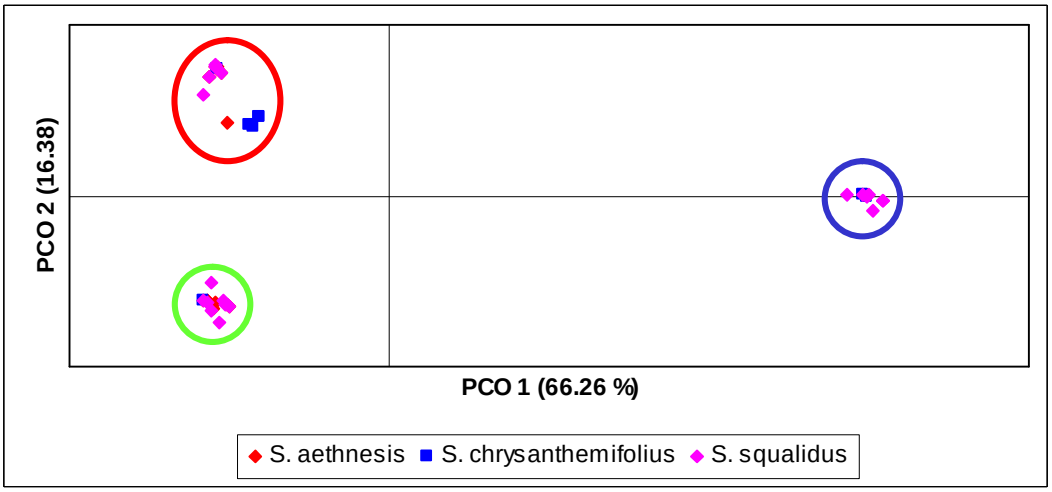


Figure 6.10: PCO plot of 46 different U61/snoR14 sequences detected in three species of *Senecio*. Within the blue encircled group, sequences from *S. aethnensis* are covered by those from *S. squalidus.*

In the NJ tree (Figure 6.11), two well supported clades (100 %) were resolved, which corresponded to the blue and the green/red encircled groups in the PCO plot (Figure 11). The mean genetic distances between clades were: A vs B = 0.028, A vs C = 0.381, B vs C = 0.334. Molecular variance among the three clades (A, B and C; Figure 6.11) was much greater (82 %) than within clades (18 %). Interestingly, sequences of a certain length are present in certain clades (Figure 6.11). For instance, clade C contained only long sequences of 121 and 131 bp, whereas clades A and B contained shorter sequences of 117 to 120 bp, and 112 to 116 bp, respectively. The size differences in sequence between the clades A and B were mainly due to a four bp deletion (at positions 32 to 36 in the alignment, Figure 6.9) present in all sequences placed in clade B. Within each of the three clades sequences from the different species were very similar and intermixed (Figure 6.10 and Figure 6.11).

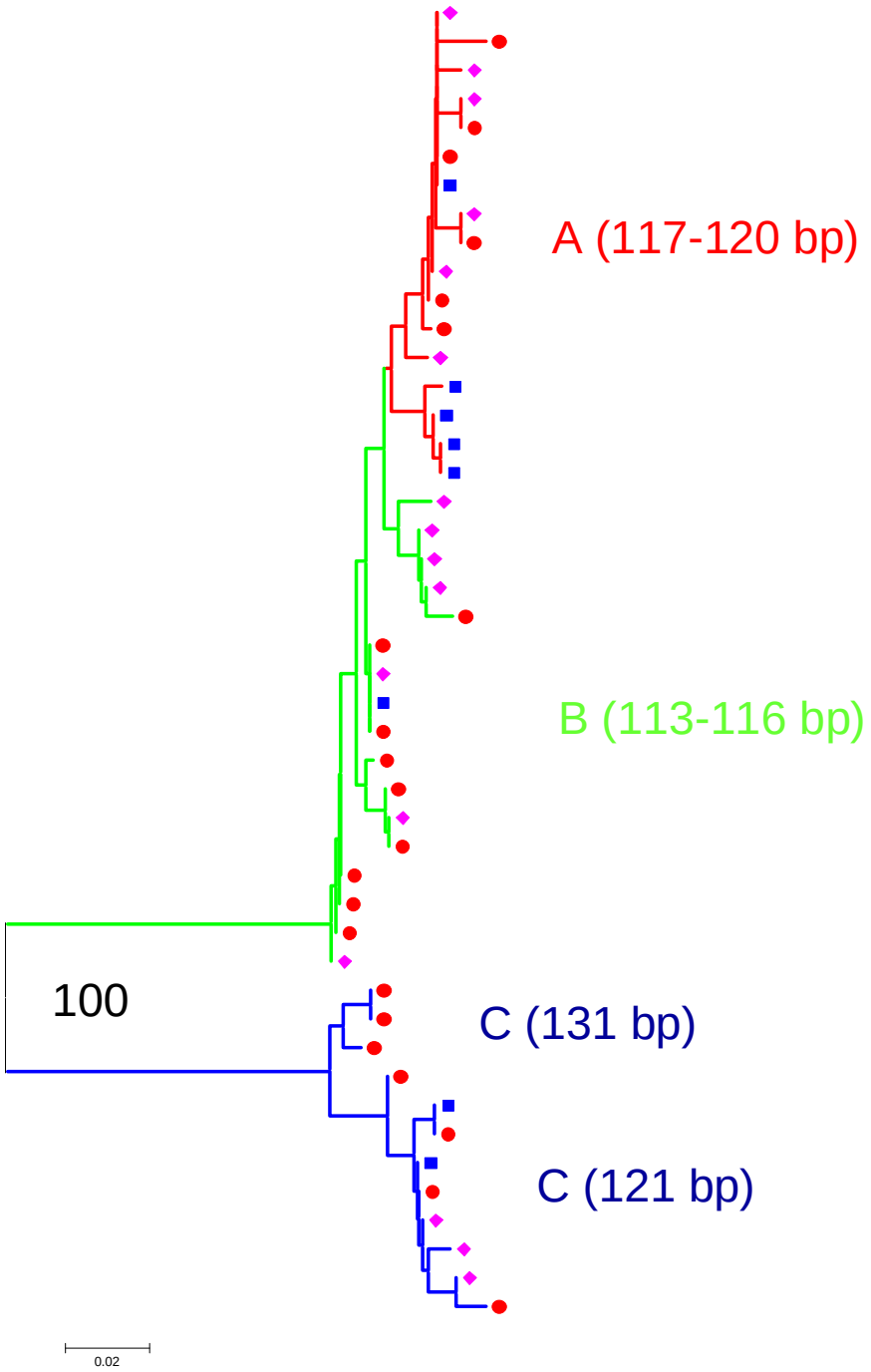


Figure 6.11: NJ tree constructed for 46 different U61F/snoR14R sequences obtained from three *Senecio* species (*S. aethnensis, S. chrysanthemifolius* and *S. squalidus*). The same colours and symbols as in the PCO plot were used to identify sequences of different species. (*S. aethnensis* = ●, *S. chrysanthemifolius* = ■ and *S. squalidus* = ♦). Bootstrap support of the two main clades (A/B vs C) is indicated and lengths of sequences are shown in brackets.

By examining both sequence and fragment analysis data it was established that each sample had at least one fragment of sequence length found in clade C (121 and 131 bp, respectively), and fragments of sequence length found in clades A and/or B. Therefore, it is feasible that clades A and B contain sequences representing different allelic variants rather than different paralogous gene copies. Thus, clade specific primers might be required to provide further insights into the different copies of each gene.

6.3.3.3 Design of clade/gene specific primers

Seven different primers were designed by 3' elongation of the original U61 universal primer (see alignment Figure 6.9). These primers consisted of one long and one short variant (e.g. U61Fc1l and U61Fc1s) for amplifying sequences assigned to each of the three clades in the NJ tree (Fig. 5.11), and an additional primer (U61Fc1_2) for amplifying sequences belonging to clades A and B. These specific primers ranged from 24 to 37 nt in length, had TMs from 48.8 to 62.2 °C, and GC contents from 28 to 44.4 % (Table 6.3).

Table 6.3: **Sequences and characteristics of specific primers designed to amplify U61/snoR14 snoRNA gene cluster sequences.** The first primer listed is the original primer sequence.

Name	Direction	Sequence (5' - 3')	Length (nt)	TM (° C)	GC (%)
U61F	forward	TACACWACCCTCTAAGAAGTTCTG	24	54	41.7
U61Fc1l		ACCCTCTAAGAAGTTCTGAGCGATTACCTTTTYTTA	36	61-62.2	36.1
U61Fc2l		ACCCTCTAAGAAGTTCTGAGCGATTACYTTTTTTTT	36	59.9-61	33.3
U61Fc3l		ACCCTCTAAGAAGTTCTGAGCAATCATTTATTATATC	37	60	32.4
U61Fc1s		GTTCTGAGCGATTACCTTTTYTTA	24	50.6-52.3	33.3
U61Fc2s		GTTCTGAGCGATTACYTTTTTTTT	24	48.8-50.6	29.2
U61Fc3s		GTTCTGAGCAATCATTTATTATATC	25	49.5	28
U61Fc1_2		ACCCTCTAAGAAGTTCTGAGCGATTAC	27	58.2	44.4

6.3.3.4 Amplification and sequence generation of specific snoRNA primer

Although long primers were designed, only short primers (s) and the additional primer (U61Fc1_2) were tested for amplification. All of these primers (Table 6.3) were successful in amplifying products from most samples examined and bands of expected sizes were obtained. For products subjected to direct sequencing, only one sequence amplified by the U61Fc3s primer was readable and, as expected, was very similar to other sequences in the clade it was assigned to (not shown).

In summary, the U61F/snoR14R sequences detected in these three *Senecio* species could be placed into three different groups, with two groups (A and B) being much more similar to each other in sequence type than either were to group C sequences. Sequences assigned to the three different groups differed in length. By combining information on sequence length variation with that obtained from the examination of fragment length profiles, it seems that only two gene cluster copies are present in these diploid species of *Senecio*, and that sequences placed in clades A and B may represent different alleles of one of these gene clusters. To examine this further, specific primers were designed and some of these successfully amplified products in some samples. However, only one good quality sequence was obtained which could be assigned clearly to its expected clade.

6.3.4 SnoR29/SnoR30

6.3.4.1 Sequence generation from original universal primers

Sixty four clones were sequenced after amplification using the SR29/SR30 universal primers, from which 53 good quality sequences were obtained. Of these, 24 represented sequences of *S. aethnensis*, 15 of *S. chrysanthemifolius*, and 14 of *S. squalidus*. After removing identical sequences within each species, 25 sequences remained for alignment (Figure 6.12). The part of the SR29 gene examined consisted of 81 bp and was separated from the SR30 gene (comprising 72 bp) by an 87 bp intergenic region. While the two gene regions were fairly well conserved, having only 11 and 15 variable sites,

respectively, the intergenic region contained 42 variable sites and was therefore less well conserved. Additionally, small gaps were detected in the SR29 and SR30 sequences, while larger gaps were present in the intergenic region. Interestingly, two sequences obtained from *S. squalidus* (squac4 and squac9) lacked box C and a small part of the antisense element of the snoR30 gene (Figure 6.12).

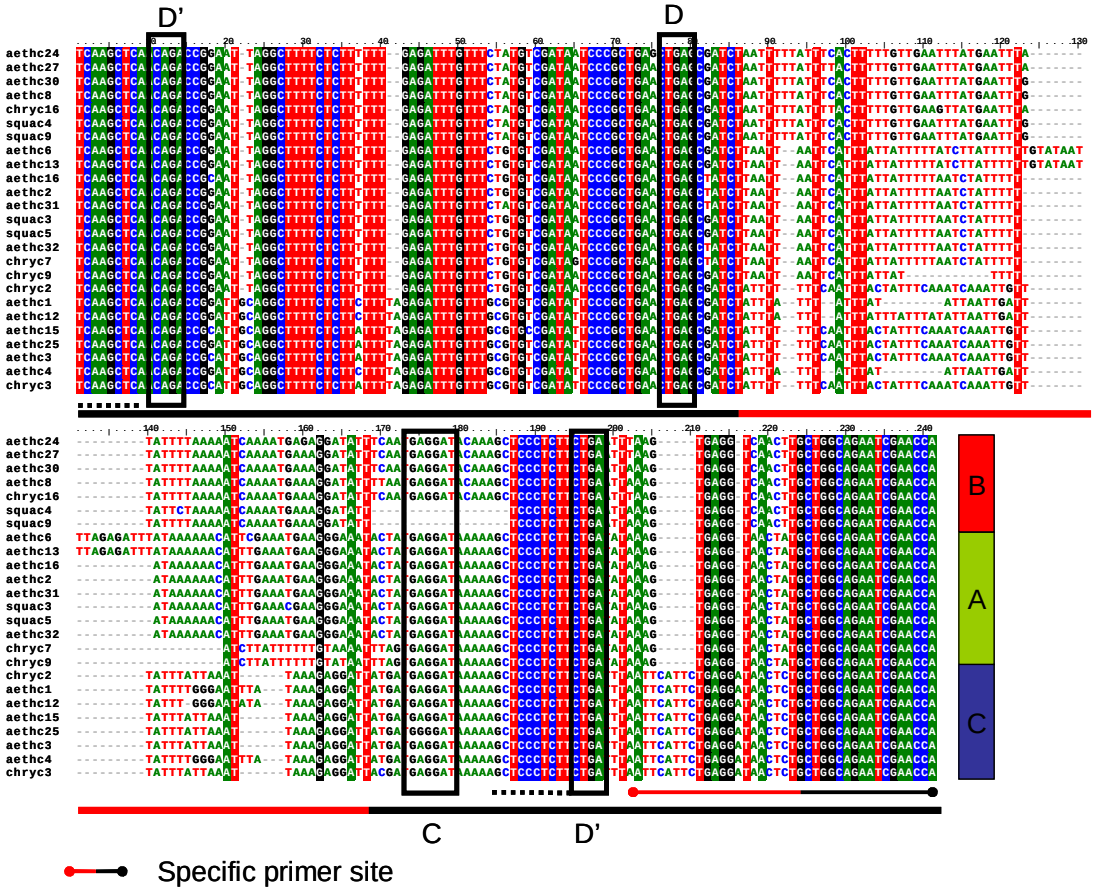


Figure 6.12: Alignment of 25 different snoR29/snoR30 sequences detected from three *Senecio* species. Conserved positions are shaded. The red part of the specific primer site symbol indicates the extension of the original primer. Coloured vertical bars indicate the clusters (A, B and C) to which the sequences belong to in the NJ tree (see below). Dotted line = antisense element; black line = gene region; red line = intergenic region. aeth = *S. aethnensis*, chry = *S. chrysanthemifolius*, squa = *S. squalidus*.

6.3.4.2 Gene copies and organization

The NJ tree (Figure 6.13) contained three well supported clades (A, B and C) with sequences of particular length represented in each. Thus, Clade B contained all sequences with lengths of 198 and 216 bp, clade A contained sequences of length 192, 203, 212 and 230 bp, partitioned into two subclades (A1 containing sequences 212/230 bp, and A2 containing sequences 192/203 bp) and clade C contained sequences of 210, 214 and 217 bp in length partitioned into two subclades, C1 and C2. The distance between clades A and B was 0.178, between clades A and C 0.264, and between clades B and C 0.228. Molecular variance among clades was much greater (76 %) than within them (24 %). Interestingly, no sequence obtained for *S. squalidus* was placed in clade C, whereas the other two clades contained sequences from all three species. Sequences from different species were intermixed within clades and consequently species were not separated according to clade (Figure 6.13).

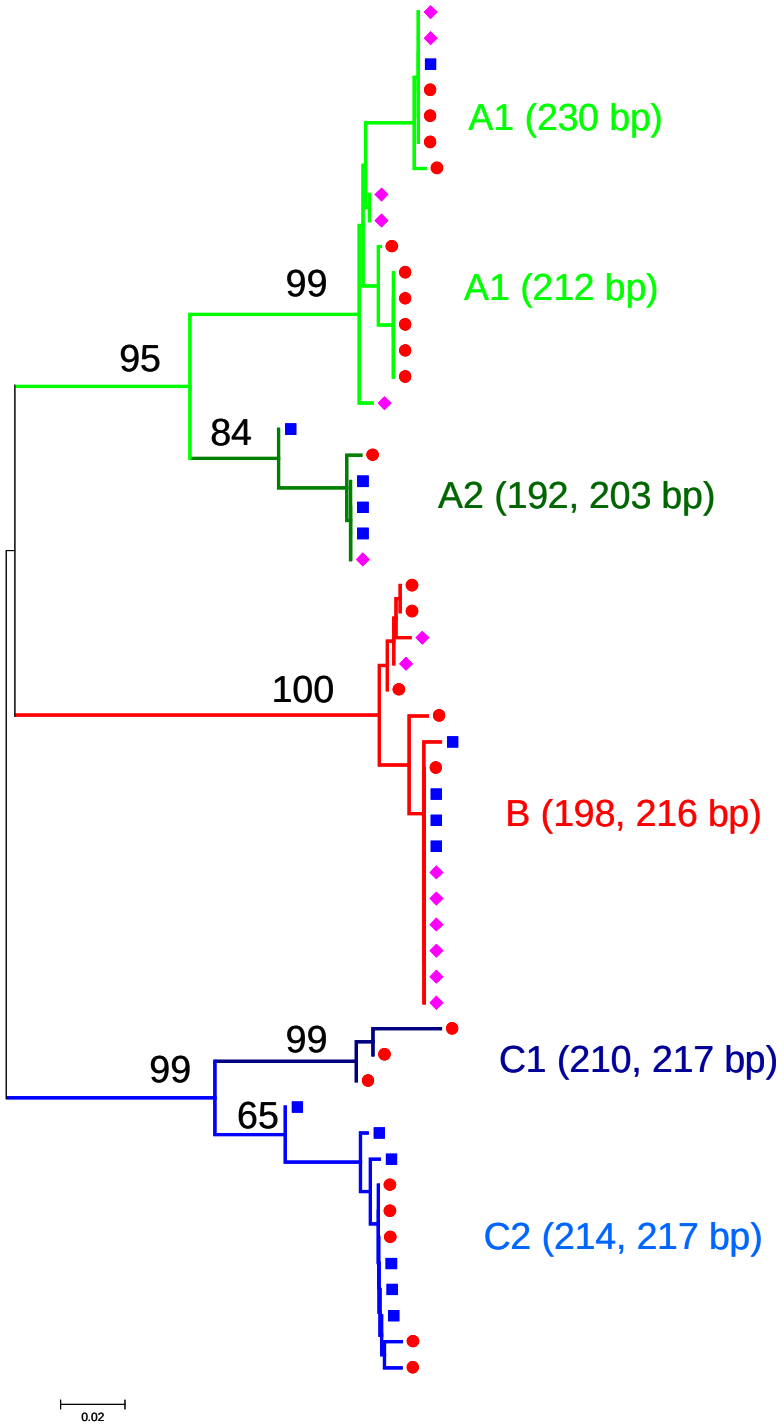


Figure 6.13: NJ tree of sequences generated by the SR29F/SR30R primer pair. Sequences of a certain length are present within a certain clade. Bootstrap values (1000 pseudoreplicates) for clades/subclades are placed above branches, while lengths of sequences are in brackets after clade or subclade labels. *S. aethnensis* = ●, *S. chrysanthemifolius* = ■ and *S. squalidus* = ♦.

The various clades identified in Figure 6.13 might contain different snoR29/snoR30 gene copies which might be discovered by the fragment profiling. Therefore, the lengths of the sequences obtained by sequence analysis were compared with the fragment profiles of all samples of *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus* produced in Chapter 4 and shown in Table 6.4.

The fragment profiles of samples contained at least three and a maximum of six fragments within the size range of sequences obtained. Clade B contained only two fragments (198 and 216 bp) and at least one fragment of 198 or 216 bp in length was found in the fragment profiles of each sample profiled (in red, Table 6.4) Surprisingly, while the 216 bp fragment was found in all *S. aethnensis*, *S. chrysanthemifolius* samples and many *S. squalidus* individuals, the 198 bp fragment was exclusively present in *S. squalidus*. Clade A contained four fragments (192, 203, 212 and 230 bp, respectively) in the sequence dataset (Figure 6.1). The 192 bp fragment was not obtained by fragment analysis, whereas one or two, and in two cases (sample 11 and 12) all three of the other size fragments were present in samples that had been subjected to fragment analysis (in green, Table 6.4). Clade C contained three fragments of 210, 214 and 217 bp length, respectively. Most samples of *S. aethnensis* and *S. chrysanthemifolius* and many individuals of *S. squalidus* contained the 217 bp fragment, whereas the 214 bp fragment was only found in six samples, and the 210 bp fragment was present in five *S. aethnensis* individuals and three of the *S. squalidus* samples examined (in blue, Table 6.4). Many of *S. squalidus* samples lacked fragments contained in clade C but showed a fragment of 209 bp which might contain clade C sequence. The remaining fragments (i.e. 200, 201 and 215 bp in lengths) could not be assigned.

When taken in combination, the sequence and fragment profile datasets appear to suggest that the well supported clades resolved in the NJ tree (i.e. clades A, B and C) contain different gene copies rather than different alleles of the SR29/SR30 genes. The development and employment of primers specific to sequences placed in these different clades specific should help to confirm or reject this proposal.

Table 6.4: **Assignation of the peaks obtained by fluorescence fragment analysis to clades obtained by sequencing.** Colours represent the clades to which the fragment might belong.

Species	Samples	fragment size (bp)											
		198	200	201	203	209	210	212	214	215	216	217	230
S. aethnensis	1							1			1	1	1
	2							1			1	1	1
	3						1	1			1		1
	4				1			1			1	1	
	5			1				1			1	1	
	6						1				1		1
	7			1			1			1	1	1	1
	8				1		1			1	1	1	1
	9			1				1	1		1	1	1
	10						1	1		1	1	1	1
S. chrysanthemifolius	11				1			1	1		1		1
	12				1			1	1		1	1	1
	13				1						1	1	1
	14				1					1	1	1	
	15										1	1	1
	17				1			1	1		1		
	18				1			1	1	1	1		
	19				1						1	1	
	20				1			1			1	1	
	21		1		1						1	1	
S. squalidus	22					1		1			1		
	23	1				1		1			1		
	24							1			1	1	
	25	1		1		1	1	1			1		
	26	1		1		1		1					
	27	1		1		1		1					
	28	1		1		1		1			1	1	
	29	1		1		1		1			1	1	
	30	1		1		1						1	1
	31	1				1					1		1
	32	1						1	1		1	1	
	33	1						1			1	1	
	34	1		1		1		1					
	35	1				1		1					
	36	1				1		1					
	37	1				1		1			1		1
	40	1			1							1	
	41	1			1	1					1		
	72	1					1	1			1		1
	73			1				1			1	1	
	74				1	1					1		
	75	1			1							1	
	76										1	1	1
	77	1			1		1				1		
	78	1			1						1	1	1

6.3.4.3 Specific primer design

Seven different primers for snoR30 consisting of one long and one short variant (e.g. SR30Rc1l and SR30Rc1s), plus one extra primer (SR30c3al), were designed to amplify sequences in each of the three clades, A, B and C (see alignment Figure 6.12). These primer sequences consisted of 19 to 29 nt, and had TMs from 49.5 to 60.1 °C, and GC contents ranging from 44.8 to 58.8 % (Table 6.5).

Table 6.5: **Sequences and characteristics of specific primers designed for the snoR29/snoR30 snoRNA gene cluster sequences.** Please note that the first primer shown is the original primer sequence.

Name	Direction	Sequence (5' - 3')	Length (nt)	TM (° C)	GC (%)
SR30F	reverse	AGCTCCCTCTTCTGA	17	49.5	58.8
SR30Rc1l		GGTTCGATTCTGCCAGCAAGTTGAC	25	59.3	52
SR30Rc2l		GGTTCGATTCTGCCAGCATAGTTAC	25	57.7	48
SR30Rc3l		GGTTCGATTCTGCCAGCAGAGTTATC	26	59.5	50
SR30Rc3al		CTGCCAGCAGAGTTATCCTCAGAATGAAT	29	60.1	44.8
SR30Rc1s		GGTTCGATTCTGCCAGCAA	19	51.1	52.6
SR30Rc2s		GGTTCGATTCTGCCAGCAT	19	51.1	52.6
SR30Rc3s		GGTTCGATTCTGCCAGCAG	19	53.2	57.9

6.3.4.4 Amplification and sequence generation by specific snoRNA primers

All snoR30 primers successfully amplified fragments of expected size (approximately 180 to 250 bp, see sequence data above) in the majority of samples. Ninety good quality sequences were obtained by cloning and sequencing products from the three species examined, i.e. *S. aethnensis* (six samples pooled), *S. chrysanthemifolius* (five samples pooled) and *S. squalidus* (four samples pooled). Twelve sequences were obtained using the c1l primer, 25 with the c2l primer, and the remaining 53 using the the c3l primer. In addition, good quality snoR30Rc1 sequences from 17 samples and also snoR30c2 sequences from 7 samples were obtained by direct sequencing. However, direct sequencing generated sequences of approximately only 160 nucleotides (one primer sequence plus adjacent nucleotides were not readable). These were aligned with sequences of the universal primer sequences. An NJ tree (not shown) generated from

these sequences was characterised by well-supported clades similar to those present in the NJ tree previously generated from sequences using universal primers (see Figure 6.13).

Due to the great differences existing between sequences representing the three different clades of the NJ tree and huge gaps within the alignment, sequences within clades were aligned and analysed separately with and without the sequences obtained by direct sequencing. As expected, all sequences produced by the snoR30Rc1 primers were placed in clade B (Figure 6.13) of the NJ tree (Figure 6.14). The genetic variation (0.012) within this clade has not changed much from the variation observed previously (i.e. analysis of universal primer sequences) and no sequences other than those of 198 and 216 bp in length were obtained using the specific primers (Figure 6.14).

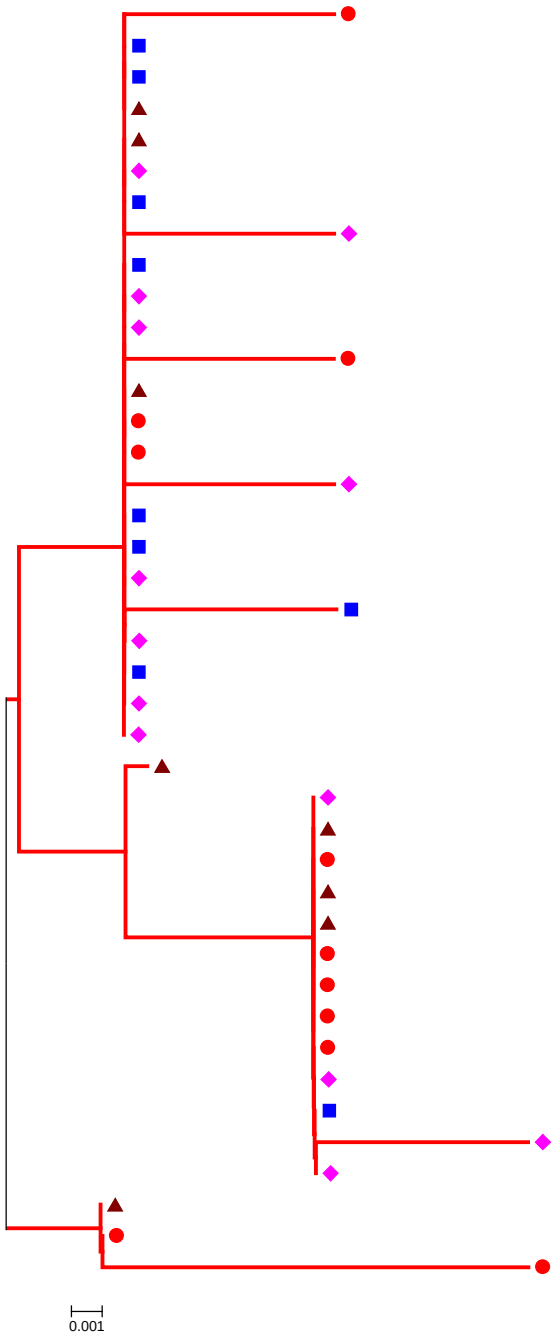


Figure 6.14: NJ tree of 40 SR29/SR30 sequences generated by primers specific to sequences placed in clade B of the NJ tree produced from sequences amplified by universal primers. The tree was produced using Maximum likelihood composition model with complete sequence elimination of gaps/missing data. *S. aethnensis* = ●, *S. chrysanthemifolius* = ■, *S. squalidus* = ♦ and *S. aethnensis* x *S. chrysanthemifolius* hybrids = ▲.

Fifty eight sequences (including seven sequences obtained by direct sequencing) were generated using primers specifically designed for sequences placed in clade A of the NJ tree produced from sequences generated by universal primers (Figure 6.13). Similarly, 52 sequences were generated using primers specifically designed for sequences placed in clade C of the universal primer NJ tree (Figure 6.13). NJ trees of both clade A and clade C sequences (Figure 6.15A, and 5.15B, respectively) generated by specific primers contained two well supported subclades. While subclade A1 (Figure 6.15A) predominantly consisted of sequences from *S. aethnensis* and *S. squalidus* (only three sequences of *S. chrysanthemifolius* were placed in this subclade), subclade A2 contained almost exclusively *S. chrysanthemifolius* sequences (only one *S. aethnensis* and one *S. squalidus* sequence were present in this subclade). In the clade C tree, subclade C2 contained no *S. squalidus* sequences (Figure 6.15). The lengths of sequences contained in the clade A tree 192, 203, 204, 206, 212 and 230 bp, and in the clade C tree were 206, 209, 210, 211, 212, 213, 214, 215 and 217 bp.

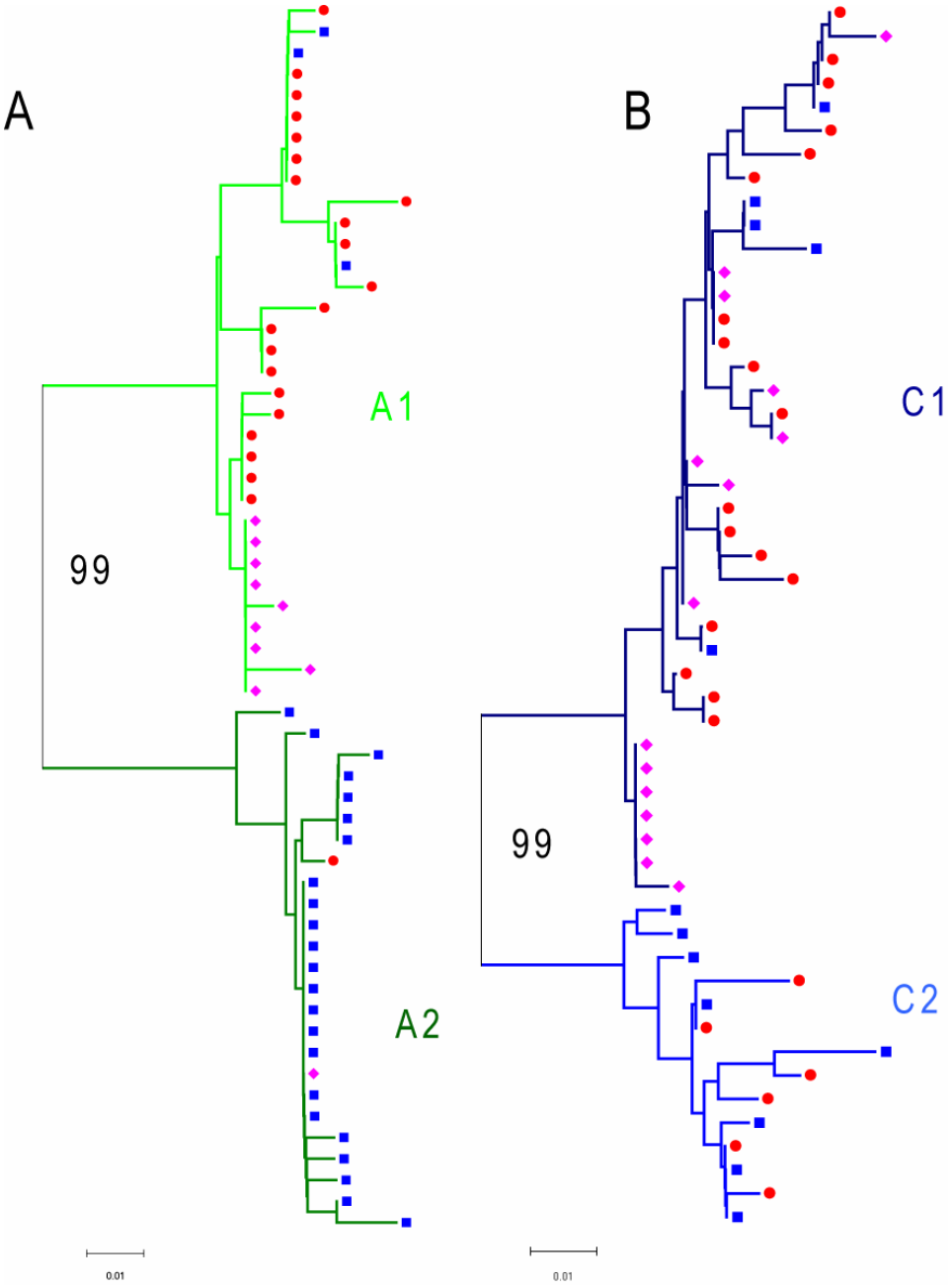


Figure 6.15: NJ trees of (A) 58 sequences and (B) 52 sequences generated by specific primers for sequences placed in clades A and C, respectively, of the NJ tree (Figure 6.13). Trees were produced using Maximum likelihood composition model with (A) pairwise and (B) complete sequence elimination of gaps data. Missing nucleotides of the shorter sequences obtained by direct sequencing were substituted with their most similar (neighbouring) sequence. *S. aethnensis* = ●, *S. chrysanthemifolius* = ■ and *S. squalidus* = ♦.

6.4 Discussion

The products of four universal snoRNA primer pairs (U33F/U51R, U14-1/U14-2, U61F/snoR14R and snoR29F/snoR30R) were examined by sequence analysis using two different approaches. While individual samples of five species (i.e. *S. aethnensis*, *S. chrysanthemifolius*, *S. squalidus*, *S. vulgaris* and *S. cambrensis*) were cloned and sequenced to examine the U33/U51 gene combinations and the homologous U14 genes (individual sample approach), the products of the U61F/snoR14R and snoR29F/snoR30R primer combination were investigated by pooling many samples of each of the three diploid species *S. aehtnensis*, *S. chrysanthemifolius* and *S. squalidus* (multiple sample approach). As shown in this study, both approaches have their advantages. The individual sample approach reveals every sequence produced by a particular primer pair and might be useful, therefore, in detecting very recent duplication events and estimating the minimum number of paralogous sequences from the number of different sequences obtained. The multiple sample approach, however, is a fast method for estimating variation in a region within and between species and to identify well differentiated paralogues. The results presented in this study confirm that these universal primers designed from *Arabidopsis thaliana* sequences amplify fragments containing the expected snoRNA gene/gene clusters. Although the sequences obtained in *Senecio* differed considerably from those in *A. thaliana*, it was shown that the structure of these regions (i.e. the lengths of genic and intergenic sequences amplified) was similar. This is in accordance with the predictions of snoRNA gene cluster organisations indicated by fragment length patterns in the previous chapter (Chapter 5). However, some of the regions amplified showed more copies than expected based on predictions from fragment lengths.

6.4.1 Duplication of snoRNA genes and gene clusters

In plants, at least one half of snoRNA genes have two to four variants arising through duplications of gene clusters on the same or different chromosome or tandem repeat duplications within clusters (Brown *et al.*, 2003a). Gene/gene cluster duplication might

lead to gene redundancy leading to relaxed evolutionary constraint and, thus, might promote the differentiation of these gene copies. Most of the regions investigated consist of gene and intergenic regions. Intergenic regions are thought to evolve neutrally and should, therefore show higher mutation rates. Thus, different paralogous regions might be identified by the differences between their sequences, especially by the intergenic region.

6.4.1.1 Recent duplication events of the U33/U51 region

The findings in this study suggest that at least three different copies of the U33/U51 regions are present in *Senecio*. The sequences were resolved into two structured clades but it was not possible to identify putative U33/U51 copies most likely due to the low sequence variation. This might be explained by recent duplication events which would not provide sufficient time for the various paralogues to accumulate mutations to be clearly differentiated and identified. The two copies of this region in *A. thaliana,* which are linked by a sequence of about 7500 bp (Brown *et al.*, 2001), also show very low differentiation (personal observation). The presence of more than one U33/U51 copy in both species might suggest a duplication event before the split of the lineages leading to *Arabidopsis* and *Senecio* and thus it could be argued that the low differentiation was caused by some degree of concerted evolution due to homogenisation mediated by unequal crossing over and/or gene conversion (Alvarez & Wendel, 2003; Baumgarten *et al.*, 2003). However, only one U33/U51 region was identified in *A. lyrata* by BLAST searches (personal observation) supporting recent duplication events in both *Arabidopsis* and *Senecio*.

6.4.1.2 Extensive gene duplication within the homologous U14 gene cluster

In *A. thaliana*, the U14 gene cluster consists of four U14 genes (Brown *et al.*, 2001), while in rice, *Oryza sativa,* as many as twenty U14 gene copies are present (identified by BLAST search, see Appendix Chapter 5). The many U14-1/U14-2 sequences identified in the *Senecio* species examined in the present study were separated into four groups by

PCO analysis, but it was not possible to distinguish between different gene copies and alleles. Based on the number of different U14-1/U14-2 sequences obtained within one sample of a species and the U14-1/U14-2 grouping pattern seen in longer sequences (i.e. sequences containing more than one gene) at least nine U14 gene copies were identified within *Senecio*, although more gene copies are likely to be found with further analysis. The various copies were shown to be grouped in clusters of up to four genes and might be the result of both tandem repeats and trans-duplication events.

6.4.1.3 Multiple snoR99 paralogues

Three sequences found within one individual of *S. squalidus* (sample 74) consisted of two U14 gene sequences separated by a long intergenic region harbouring the snoR99 gene. This might suggest a duplication of the complete or part of the U14 gene cluster to the same or different chromosome. However, as these sequences were identical except for some minor differences within the 3' U14 copy, it is more likely that the snoR99 gene was tandemly duplicated after a U14 gene copy with a 3' adjacent snoR99 gene was aligned to an identical U14 gene copy lacking this adjacent gene.

6.4.1.4 High differentiation between putative U61/snoR14 and snoR29/snoR30 paralogues

The U61/snoR14 sequence is a single copy region in most species including *A. thaliana* (Brown *et al.*, 2001) and *O. sativa* (Chen *et al.*, 2003). In *A. thaliana* a second copy of the snoR14 gene is also present (Brown *et al.*, 2001). In *Senecio*, PCO analysis of sequence variation distinguished three different groups for this region, two of which were highly supported in a NJ tree. These different groups differed in sequence length and, when compared to the data obtained from fragment analysis (see Chapter 4), indicated the presence of two copies of the U61/snoR14 region in *Senecio*. These two putative copies showed a high degree of differentiation in their intergenic region but were highly conserved in their gene regions, suggesting a relatively ancient duplication event coupled with high evolutionary constraint in the gene regions.

Sequences of the snoR29/snoR30 region obtained for *Senecio* were placed into three well supported clades (A, B and C) in a NJ tree, suggesting that three copies of this region are present in the genus which is in accordance with the findings of the fragment analysis reported in Chapter 5. In *A. thaliana* this region is represented by only one copy (Brown *et al.*, 2001), while three copies have been identified in rice (Chen *et al.*, 2003). The three different *Senecio* sequences differed in both intergenic and gene regions. These differences would suggest a duplication resulted first in the formation and subsequent divergence of the A/B and C copies, followed by a more recent duplication causing the formation of the A and B copies. In the initial analysis using universal primers each of these putative gene copies appeared to be distinguished according to length. However, a more detailed analysis using specific primer pairs showed overlap in the sizes of some A and C sequences and, thus, caution is required when attempting to identify these different copies using only fragment length data.

In the NJ trees derived from sequences using specific primers, both the A and C clades contained two highly supported subclades probably indicating the presence of additional snoR29/snoR30 copies. Different gene copies would result in the amplification of sequences of both subclades within each individual and, thus, an almost equal number of each species' sequences would be expected within each subclade (note that the same result would also be expected if the subclades represent different alleles and would be equally frequent in each species). Because of the predominance of a certain species within one subclade and almost entire lack of this species within the other subclade, extra snoR29/snoR30 copies within the clade A can be safely ruled out. Similarly, snoR29/snoR30 subclade C2 lacked *S. squalidus'* sequences and, therefore, additional copies within clade C are highly unlikely.

6.4.2 Sequence variation between *Senecio* species

Although it was not possible to identify the various putative gene copies of the U33/U51 region in *Senecio*, the sequences obtained from *S. vulgaris* differed considerably from all species except *S. cambrensis* which possessed identical sequences. These results reflect both the distant relationship between *S. vulgaris* and the diploid species *S. aethnensis*, *S.*

chrysanthemifolius and *S. squalidus* (Comes & Abbott, 2001; Coleman *et al.*, 2003; Pelser *et al.*, 2007; Milton, 2009) and the parent-hybrid relationship between *S. vulgaris* and the allohexaploid, *S. cambrensis* (Abbott *et al.*, 1992; Harris & Ingram, 1992b; Abbott & Lowe, 2004). *Senecio cambrensis* also possessed identical/highly similar sequences to *S. squalidus* its other parent.

S. squalidus is the hybrid of *S. aethnensis* and *S. chrysanthemifolius* (Abbott *et al.*, 2000; Abbott *et al.*, 2002; James & Abbott, 2005) and should therefore contain sequences identical or highly similar to those present in its parents. Most sequences identified in *S. squalidus* were in fact identical to those in *S. chrysanthemifolius* but differed from those found in *S. aethnensis*. This would suggest a higher proportion of the *S. squalidus* genome is derived from *S. chrysanthemifolius* than from *S. aethnensis*, which is in accordance with previous results based on surveys of RAPD/ISSR markers in all three species (James & Abbott, 2005).

Identical U14 sequences were found to be present in distantly related species such as *S. vulgaris* and the diploid species *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus*. This high degree of conservation across species might indicate purifying selection on the function of these genes (Schmitz *et al.*, 2008). At the same time, the high variability of paralogues might reflect extensive duplication and subsequent mutations due to gene redundancy (Brown *et al.*, 2001; Brown *et al.*, 2003a).

Two putative U61/snoR14 and three putative snoR29/snoR30 paralogous regions (i.e. well supported clades) were examined for their ability to separate the three diploid species *S. aethnensis*, *S. chrysanthemifolius* and *S. squalidus*. For each region sequences obtained from each species were intermixed between taxa and, hence, these species were never clearly separated. These findings are not surprising given the close relationship between these species and clearly demonstrate that the within species variation is higher than the variation between these species. However, the species were separated, albeit with various degrees of intermixing between the taxa. Interestingly, the degree of species intermixing varied greatly between different paralogous copies, especially between snoR29/snoR30 copies. While the snoR29/snoR30 clade A copy was almost able to separate *S. aethnensis* and *S. chrysanthemifolius* into two different subclades, the clade C copy showed more intermingling between these species and the sequences of the clade B

copy were highly conserved between these species. The former two copies appear to differ in the frequencies of certain sequences within these species, whereas the latter one might be subjected to purifying selection (Schmitz *et al.*, 2008) which might act on the region this copy is placed in. The putative U61/snoR14 copies also tended to share identical sequences between species, but some sequences, although highly similar between species, appeared to be unique for certain taxa. Therefore, it is feasible that the species examined show differences in the frequencies of their U61/snoR14 sequences.

In addition to the nucleotide polymorphisms, the sequences of each putative paralogous region (i.e. putative U61/snoR14 and snoR29/snoR30 copies) showed variation in their length due to indels. The indels were shown to be mostly (sub)clade specific emphasising the differences of these (sub)clades. Differences in sequence length can be easily identified by fragment analysis and might subsequently be scored as co-dominant markers and useful for differentiation of various species. As shown in this chapter, the fragment profiles of the universal snoR29/snoR30 and U61/snoR14 primer combinations revealed that certain sequences were highly frequent in one species and rare or absent within another species. For example, the putative U61/snoR14 clade C copy consisted of sequences with lengths of 121 and 131 bp. The latter was highly frequent in *S. aethnensis* but absent in the other species and, thus, shows some degree of differentiation.

6.4.3 Sequence variation of snoRNA genes and functional evolution

As shown in this study, all of the investigated regions are present in at least two copies which show various degrees of sequence variation within their genic regions. While most of the variation might not change the function of the snoRNA genes, some mutation, especially within the conserved regions (i.e. the boxes C and D and the antisense element) might alter function. The box C/D motif is essential for the formation, stability and function of the small nucleolar ribonucleoprotein particle (snoRNP) (Samarsky *et al.*, 1998), whereas the antisense element basepairs with the rRNA (Makarova & Kramerov, 2007). Thus, the snoRNA modification ability might be affected by mutations in these regions. Interestingly, the snoR14 gene, although highly conserved, showed some

variation within its box C sequences and antisense element. These variations were present in each gene cluster and could result from allelic variation and/or relaxed evolutionary constraints due to gene redundancy (Brown *et al.*, 2003a). The variation observed within the antisense element might also stem from compensatory mutations to maintain the rRNA-snoRNA duplex after a putative mutation within the complementary rRNA sequence (Chen *et al.*, 2003).

Similarly, one sequence of both *S. squalidus* samples and one *S. cambrensis* sample differed in the C box sequences of the U51 snoRNA gene suggesting that the mutation arose in *S. squalidus* and was subsequently inherited by *S. cambrensis*. However, this difference in box C sequence might also represent allelic variation due to gene redundancy and the lack of this sequence in other species might be explained by the low number of samples investigated.

Some sequences of the putative snoR29/snoR30 clade B copy lacked the box C element of the snoR30 gene resulting in a non-functional sequence. Interestingly, this sequence was identified in most *S. squalidus* samples by the universal primer fragment profiles, but was absent in samples of *S. aethnensis*, *S. chrysanthemifolius* and also their natural hybrids. Furthermore, some of the *S. squalidus* samples were homozygous for the non-functional allele indicating that the individual function of most snoRNAs might be non-essential (Brown *et al.*, 2003a) and, thus, the non-functional snoR30 sequences might be tolerated. As this sequence was unique to and highly frequent in *S. squalidus* it has probably arisen and spread within this species during the early stages of colonisation of the United Kingdom and Ireland. It might also be possible that this sequence is rare in *S. aethnensis* and/or *S. chrysanthemifolius* and therefore was not present in the few samples examined. In this case the non-functional snoR30 allele might have been inherited by the hybrid material introduced to Great Britain and subsequently spread during colonisation.

6.4.4 snoRNA markers and their application in DNA barcoding and phylogenetic studies

All of the regions used for sequence analysis were shown to be present in more than one copy in *Senecio*. The paralogous copies of some regions (i.e. U33/U51 and U14 cluster)

are unsuitable for their use in DNA barcoding and phylogenetic studies due to problems stemming from an inability to distinguish between orthologous and paralogous genes. Although the copies of other regions (i.e. U61/snoR14 and snoR29/snoR30) could be identified and are most likely single copy regions, they are also not applicable for DNA barcoding because copy specific primer is necessary for amplification. Additionally, these regions were shown to be present in more copies in *Senecio* relative to *Arabidopsis*. Differences in the copy numbers between species, which are usually distantly related, might lead to difficulties in the identification of orthologous copies, and therefore are another reason to exclude these regions from DNA barcoding. However, snoRNA genes and gene clusters might be highly useful for phylogenetic studies of species groups, genera and families because the various paralogous copies are highly likely to be present in all species and the modified specific primer might be universally applied across the group of interest.

Although the putative single copy regions (i.e. U61/snoR14 and snoR29/snoR30 copies) were not able to clearly separate very closely related species, they showed some degree of differentiation between these species. By combining several regions it might be possible to clearly discriminate between closely related species and reveal reticulate evolution. Therefore, some of these snoRNA markers are applicable in phylogenetic studies and might serve as additional DNA barcodes to delimit taxa in difficult plant groups. Furthermore, each of these regions displayed some degree of sequence length variation and various alleles might be identified by fragment analysis and subsequently used as co-dominant markers for studies based on allele frequencies (e.g. population genetics). However, before these regions can be used for future studies, their single copy status has to be confirmed. Further investigation using specific primers and individual sample approaches might help to achieve this.

Chapter 7: General discussion

7.1 Development of a snoRNA marker system for phylogenetic studies and DNA barcoding

SnoRNA genes and gene clusters should have great potential for use in phylogenetic studies and DNA barcoding of species because they (i) are thought to evolve faster than protein coding genes, (ii) are scattered across the genomes of species, (iii) are single or low copy regions present across plant families, (iv) comprise short sequences, and (v) provide highly conserved regions for annealing of primers. The major objective of the research reported in this thesis was to investigate this potential and to test, therefore, whether snoRNA genes and gene clusters do indeed have advantages over other molecular marker systems in phylogenetic studies of closely related species and in DNA barcoding.

Universal primers for amplifying snoRNA genes were designed using snoRNA gene sequences identified in *Arabidopsis thaliana* which were aligned to homologues from other plant species found by BLAST search. As most snoRNA genes contained only one putative primer site, at least one other snoRNA gene within the same gene cluster was usually required to provide a second primer site. The majority of snoRNA gene clusters showed highly conserved organization allowing the combination of various primer pairs for successful amplification. Primers were characterized according to sequence, and possible primer combinations were tested virtually. The success of this approach was demonstrated using five different snoRNA gene clusters (Chapter 3).

Most primer pairs designed using *A. thaliana* sequences successfully amplified snoRNA genes and gene clusters in the majority of *Senecio* species tested using a standardized protocol. The fragments produced were scored as dominant markers and often showed variation between and within species. Thus, by examining the variation pattern over a few datasets generated by different primers it was possible to delimit species and detect reticulate evolution (Chapter 4). It was concluded that snoRNA gene/gene cluster fragment length polymorphisms (SRFLPs) can be used as a universal

marker system for studying phylogenetic relationships between closely related species in the genus *Senecio*, and other plant genera by extension.

Although SRLPs (snoRNA gene/gene cluster fragment length polymorphisms) were shown to have high potential for resolving phylogenetic relationships between closely related species, the full potential of this marker system has yet to be investigated. In Chapter 4 the results were based on the amplification products of only a few primer pairs, whereas in *A. thaliana* there are more than 175 snoRNA genes organized into at least 49 clusters scattered across the genome (Brown *et al.*, 2003a; Chen & Wu, 2009). Thus there is a very large pool of further possible markers to choose from, which should make it possible to select many more suitable and informative snoRNA regions for use in future phylogenetic applications.

Most of the snoRNA genes and gene clusters investigated in *Senecio* were found to be present in more than one copy and it was for this reason that fragments were scored as dominant markers. However, for some snoRNA genes, sequence variation between orthologous and paralogous copies was shown to be useful for isolating single copy regions, which might subsequently be developed as co-dominant markers (Chapter 6). The sequences of these putative single copy regions often varied considerably in length making them easy to distinguish in phylogenetic analyses.

The sequences of some snoRNA regions, especially putative single copy regions (i.e. paralogous copies of various clusters), were investigated for their ability to distinguish between closely related species of *Senecio* and, therefore, for their potential use in DNA barcoding. The analysis showed that none of these regions discriminated between species and consequently they were considered unsuitable for DNA barcoding. However, as already mentioned there are many snoRNA genes and gene clusters yet to be investigated and, thus, it is feasible that some of these will be useful in distinguishing between closely related species in future studies. That said, the majority of snoRNA genes and gene clusters are most likely found in more than one copy across plant families (Brown *et al.*, 2003a; Chen *et al.*, 2003; Li *et al.*, 2007; Chen *et al.*, 2008; Chen & Wu, 2009), which excludes them from DNA barcoding because (i) primer modification for specific copies would be necessary and hence universal primers could not be used in the analysis, and (ii) it might be difficult to decide which of the copies should be used in

DNA barcoding. For example, in this study, three and two putative copies of snoR29/snoR30 and U61/snoR14, respectively, were identified in *Senecio* but only a single copy is present in *Arabidopsis thaliana.* This raises the question as to which of the variants in *Senecio* are orthologous to the *A. thaliana* sequences.

Although most snoRNA gene and gene cluster sequences might not be suitable for plant DNA barcoding, they might be highly useful for phylogenetic reconstruction of species groups, genera and families. For example, the three putative snoR29/snoR30 paralogues might be present in all Asteraceae and could, therefore, be amplified across the family by specific primers designed for each copy. In this case three putative single copy regions, amplified by universal Asteraceae primers, could help improve phylogenetic resolution at all taxonomic levels within the family.

7.2 Characterisation and evolution of snoRNA genes and gene clusters

The other goals of the research presented in this study were to characterize snoRNA genes and gene clusters in *Senecio* and to investigate their evolution.

The multiple isoforms of many snoRNA genes found in plants are the result of cis- and trans-duplications of gene clusters and tandem repeats within a cluster (Barneche *et al.*, 2001). From the results presented in Chapters 5 and 6, it can be concluded that most of the snoRNA genes and gene clusters examined in *Senecio* were present in more copies relative to those present in *Arabidopsis thaliana* and have most likely arisen through one or more of the proposed mechanisms of cis-, trans- and tandem duplication. For example, tandem repeats might have played a major role in the duplication of U14 genes, whereas cis- and trans-duplications might have been the predominant process for the duplication of snoR29/snoR30 and U61/snoR14 clusters, respectively.

Rearrangements at the level of snoRNA gene cluster organization could lead to the loss of entire genes but has most likely contributed to the high diversity of plant snoRNAs due to unequal crossing over and gene conversions (Barneche *et al.*, 2001; Brown *et al.*, 2001; Qu *et al.*, 2001). Comparisons between different plant species showed that the gene order of some clusters is highly conserved whereas other clusters

are mixed and dispersed (Brown *et al.*, 2001; Brown *et al.*, 2003a; Chen *et al.*, 2003). As shown in Chapter 5 and 6, most of the gene clusters characterized in *Senecio* appear to be highly similar in organization to those in *A. thaliana*. However, some clusters appear to differ as a result of tandem repeats, inversions and gene losses.

The high level of snoRNA variants in plants leads to gene redundancy which thus provides an opportunity for the accumulation of mutations. However, despite the occurrence of large sequence differences between paralogues, the functionality of snoRNA isoforms is likely to remain unchanged, as has been shown to be the case for many snoRNA isoforms found in *Oryza sativa* and *Arabidopsis thaliana* (Brown *et al.*, 2003a; Brown *et al.*, 2003b; Chen *et al.*, 2003; Chen *et al.*, 2008). That said, variation within the box C/D motif and the rRNA complementary region might modify the functionality of some snoRNAs (Samarsky *et al.*, 1998; Makarova & Kramerov, 2007). Mutations in the box C/D motif could lead to very non-canonical sequences and subsequently result in non-functional pseudogenes (Li *et al.*, 2007), while alterations in rRNA antisense elements could produce novel modification sites (Brown *et al.*, 2003a). In the present study, mutations in the box C/D motif were noted from the alignment of three regions (U33/U51, U61/snoR14 and snoR29/snoR30, respectively). Furthermore, sequences of one putative snoR29/snoR30 paralogue showed a deletion of the entire snoR30 box C element leading to a non-functional pseudogene. This appeared to be present in a homozygous state in some *S. squalidus* samples emphasising the functional non-essentiality of individual snoRNAs. Only one snoRNA gene showed variation in its rRNA antisense element which might suggest a non-functional gene, a new one with a novel modification site or a compensatory mutation to maintain the rRNA-snoRNA interaction and hence rRNA modification.

In summary, although only a few snoRNA genes and gene clusters were investigated in the present research, it was shown that most were present in multiple copies. The sequences of some snoRNA gene copies differ considerably and might under selection lead to the occurrence of new functional snoRNA genes within species. Therefore, snoRNA genes might provide an excellent marker system for studying gene evolution and by comparing snoRNA gene clusters between various species it might be

possible to investigate the origin and evolution of gene clusters. Sequencing of entire gene clusters using flanking primers will be necessary for such detailed analysis.

7.3 Concluding remarks and future directions

Although snoRNA genes and gene clusters were not shown to be useful for DNA barcoding in *Senecio*, they appear to hold great promise for use in phylogenetic studies and the investigation of gene and gene cluster evolution. The high number of snoRNA gene and gene clusters spread across the entire genomes of plants provides a large pool of potential markers for future investigation. However, before the full potential of snoRNA markers can be appreciated many more snoRNA genes and gene clusters will have to be investigated. The next steps in developing this system in *Senecio* should focus on sequencing each of the remaining gene clusters investigated in this study and to clearly confirm the single copy status of the putative paralogous identified. Other snoRNA genes and gene clusters identified in *Arabidopsis thaliana* might be investigated in *Senecio* by using the guidelines shown in this thesis. EST libraries for *S. aethnensis, S. chrysanthemifolius, S. squalidus*, *S. vulgaris* and *S. cambrensis* are now available (Simon Hiscock, University of Bristol, UK, personal communication) and an EST database for *S. madagascariensis* is in the process of being generated (Andrew Lowe, University of Adelaide, Australia, personal communication). These EST libraries should be screened for additional snoRNAs using bioinformatic tools like the snoRNA platform (Chen *et al.*, 2003; Huang *et al.*, 2007; Chen *et al.*, 2008), snoScan (Lowe & Eddy, 1999) and SnoGPS (Schattner *et al.*, 2004), while comparative analysis, secondary structure prediction and the identification of modification sites could be used for further characterization of these genes (e.g. Chen *et al.*, 2008; Chen & Wu, 2009). These approaches will likely generate many more snoRNA markers for future use in investigations focused on elucidating patterns of gene, gene cluster and genome evolution, and phylogenetic reconstruction.

References

Abbott RJ (1992) Plant invasions, interspecific hybridization and the evolution of new plant taxa. *Trends in Ecology & Evolution* **7**, 401-405.

Abbott RJ, Ashton PA, Forbes DG (1992) Introgressive origin of the radiate groundsel, *Senecio vulgaris* L var *hibernicus* Syme - aat-3 evidence. *Heredity* **68**, 425-435.

Abbott RJ, Brennan AC, James K, *et al.* (2009) Recent hybrid origin and invasion of the British Isles by a self-incompatible species, Oxford ragwort (*Senecio squalidus* L., Asteraceae). *Biological Invasions* **11**, 1145-1158.

Abbott RJ, Ireland HE, Rogers HJ (2007) Population decline despite high genetic diversity in the new allopolyploid species *Senecio cambrensis* (Asteraceae). *Molecular Ecology* **16**, 1023-1033.

Abbott RJ, James JK, Forbes DG, Comes HP (2002) Hybrid origin of the Oxford Ragwort, *Senecio squalidus* L: Morphological and allozyme differences between *S. squalidus* and *S. rupestris* Waldst. and Kit. *Watsonia* **24**, 17-29.

Abbott RJ, James JK, Irwin JA, Comes HP (2000) Hybrid origin of the Oxford Ragwort, *Senecio squalidus* L. *Watsonia* **23**, 123-138.

Abbott RJ, Lowe AJ (2004) Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. *Biological Journal of the Linnean Society* **82**, 467-474.

Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* **29**, 380-395.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Alvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**, 417-434.

Ashton PA, Abbott RJ (1992) Multiple origins and genetic diversity in the newly arisen allopolyploid species, *Senecio cambrensis* Rosser (Compositae). *Heredity* **68**, 25-32.

Atzorn V, Fragapane P, Kiss T (2004) U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Molecular and Cellular Biology* **24**, 1769-1778.

Bachellerie JP, Cavaille J (1997) Guiding ribose methylation of rRNA. *Trends in Biochemical Sciences* **22**, 257-261.

Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. *Biochimie* **84**, 775-790.

Baldwin BG, Sanderson MJ, Porter JM, *et al.* (1995) The ITS region of nuclear ribosomal DNA - a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* **82**, 247-277.

Balloux F, Goudet J (2002) Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* **11**, 771-783.

Bangham CRM (1991) The polymerase chain reaction: Getting started. In: *Protocols in human molecular genetics* (ed. Mathew C), pp. 1-8. Humana Press, Clifton, NJ.

References

Bao Y, Wendel J, Ge S (2010) Multiple patterns of rDNA evolution following polyploidy in *Oryza*. *Molecular Phylogenetics and Evolution* **55**, 136-142.

Barkman TJ, Simpson BB (2002) Hybrid origin and parentage of D*endrochilum acuiferum* (Orchidaceae) inferred in a phylogenetic context using nuclear and plastid DNA sequence data. *Systematic Botany* **27**, 209-220.

Barneche F, Gaspin C, Guyot R, Echeverria M (2001) Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: Extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2 '-O-methylation sites. *Journal of Molecular Biology* **311**, 57-73.

Baumgarten A, Cannon S, Spangler R, May G (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics* **165**, 309-319.

Bayly M, Ladiges P (2007) Divergent paralogues of ribosomal DNA in eucalypts (Myrtaceae). *Molecular Phylogenetics and Evolution* **44**, 346-356.

Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciencies* **359**, 669-679.

Bonaldo MDF, Lennon G, Soares MB (1996) Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Research* **6**, 791-806.

Bonin A, Bellemain E, Eidesen PB, *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261-3273.

Borovjagin AV, Gerbi SA (1999) U3 small nucleolar RNA is essential for cleavage at sites 1, 2 and 3 in pre-rRNA and determines which rRNA processing pathway is taken in *Xenopus* oocytes. *Journal of Molecular Biology* **286**, 1347-1363.

References

Borovjagin AV, Gerbi SA (2000) The spacing between functional cis-elements of U3 snoRNA is critical for rRNA processing. *Journal of Molecular Biology* **300**, 57-74.

Borovjagin AV, Gerbi SA (2001) *Xenopus* U3 snoRNA GAC-box A ' and box A sequences play distinct functional roles in rRNA processing. *Molecular and Cellular Biology* **21**, 6210-6221.

Borovjagin AV, Gerbi SA (2005) An evolutionary intra-molecular shift in the preferred U3 snoRNA binding site on pre-ribosomal RNA. *Nucleic Acids Research* **33**, 4995-5005.

Bortolin ML, Ganot P, Kiss T (1999) Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. *Embo Journal* **18**, 457-469.

Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4661-4665.

Brown JWS, Clark GP, Leader DJ, Simpson CG, Lowe T (2001) Multiple snoRNA gene clusters from *Arabidopsis*. *Rna-a Publication of the Rna Society* **7**, 1817-1832.

Brown JWS, Echeverria M, Qu LH (2003a) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends in Plant Science* **8**, 42-49.

Brown JWS, Echeverria M, Qu LH, *et al.* (2003b) Plant snoRNA database. *Nucleic Acids Research* **31**, 432-435.

References

Brown JWS, Marshall DF, Echeverria M (2008) Intronic noncoding RNAs and splicing. *Trends in Plant Science* **13**, 335-342.

Brown JWS, Shaw PJ (1998) Small nucleolar RNAs and pre-rRNA processing in plants. *Plant Cell* **10**, 649-657.

Bungard R (2004) Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. *BioEssays* **26**, 235-247.

Busch H, Reddy R, Rothblum L, Choi YC (1982) SnRNAs, snRNPs, and RNA processing. *Annual Review of Biochemistry* **51**, 617-654.

Cahill NM, Friend K, Speckmann W, *et al.* (2002) Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box C/D snoRNP. *Embo Journal* **21**, 3816-3828.

Campbell CS, Wojciechowski MF, Baldwin BG, Alice LA, Donoghue MJ (1997) Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier agamic* complex (Rosaceae). *Molecular Biology and Evolution* **14**, 81-90.

Cárdenas-Flores A, Draye X, Bivort C, Cranenbrouck S, Declerck S (2010) Impact of multispores in vitro subcultivation of *Glomus* sp. MUCL 43194 (DAOM 197198) on vegetative compatibility and genetic diversity detected by AFLP. *Mycorrhiza* **20**, 415-425.

Chanfreau G, Legrain P, Jacquier A (1998) Yeast RNase III as a key processing enzyme in small nucleolar RNAs metabolism. *Journal of Molecular Biology* **284**, 975-988.

References

Chapman MA, Chang J, Weisman D, Kesseli RV, Burke JM (2007) Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). *Theoretical and Applied Genetics* **115**, 747-755.

Chase MW, Cowan RS, Hollingsworth PM, *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon* **56**, 295-299.

Chase MW, Fay, M.F. (2001) Ancient flowering plants: DNA sequences and aniosperm classification. *Genome Biology* **2(4)**, 1012.1011-1012.1014.

Chase MW, Salamin N, Wilkinson M, *et al.* (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **360**, 1889-1895.

Chen CL, Chen CJ, Vallon O, *et al.* (2008) Genomewide analysis of box C/D and box H/ACA snoRNAs in *Chlamydomonas reinhardtii* reveals an extensive organization into intronic gene clusters. *Genetics* **179**, 21-30.

Chen CL, Liang D, Zhou H, *et al.* (2003) The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Research* **31**, 2601-2613.

Chen H-M, Wu S-H (2009) Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in *Arabidopsis*. *Nucleic Acid Research* **37**, 1-12.

Chen MS, Goswami PC, Laszlo A (2002) Differential accumulation of U14 snoRNA and hsc70 mRNA in chinese hamster cells after exposure to various stress conditions. *Cell Stress & Chaperones* **7**, 65-72.

References

Coleman M, Abbott RJ (2003) Possible causes of morphological variation in an endemic Moroccan groundsel (*Senecio leucanthemifolius* var. *casablancae*): evidence from chloroplast DNA and random amplified polymorphic DNA markers. *Molecular Ecology* **12**, 423-434.

Coleman M, Liston A, Kadereit JW, Abbott RJ (2003) Repeat intercontinental dispersal and Pleistocene speciation in disjunct Mediterranean and desert *Senecio* (Asteraceae). *American Journal of Botany* **90**, 1446-1454.

Comes HP, Abbott RJ (1999) Population genetic structure and gene flow across arid versus mesic environments: A comparative study of two parapatric *Senecio* species from the Near East. *Evolution* **53**, 36-54.

Comes HP, Abbott RJ (2001) Molecular phylogeography, reticulation, and lineage sorting in Mediterranean *Senecio* sect. *Senecio* (Asteraceae). *Evolution* **55**, 1943-1962.

Darzacq X, Jady BE, Verheggen C, *et al.* (2002) Cajal body-specific small nuclear RNAs: a novel class of 2 '-O-methylation and pseudouridylation guide RNAs. *Embo Journal* **21**, 2746-2756.

Darzacq X, Kiss T (2000) Processing of intron-encoded box C/D small nucleolar RNAs lacking a 5 ',3 '-terminal stem structure. *Molecular and Cellular Biology* **20**, 4522-4531.

Dennis PP, Omer A (2005) Small non-coding RNAs in Archaea. *Current Opinion in Microbiology* **8**, 685-694.

Dennis PP, Omer A, Lowe T (2001) A guided tour: small RNA function in Archaea. *Molecular Microbiology* **40**, 509-519.

References

dePamphilis C, Palmer J (1990) Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature* **348**, 337-339.

Doyle JJ, Doyle JS (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulltetin* **9**, 11-15.

Dunbar DA, Baserga SJ (1998) The U14 snoRNA is required for 5 '-O-methylation of the pre-18S rRNA in *Xenopus* oocytes. *Rna-a Publication of the Rna Society* **4**, 195-204.

Ehrich D (2006) AFLPDAT: a collection of R functions for convenient handling of AFLP data. *Molecular Ecology Notes* **6**, 603-604.

Ehrich D, Gaudeul M, Assefa A, *et al.* (2007) Genetic consequences of Pleistocene range shifts: contrast between the Arctic, the Alps and the East African mountains. *Molecular Ecology* **16**, 2542-2559.

Ero R, Peil L, Liiv A, Remme J (2008) Identification of pseudouridine methyltransferase in *Escherichia coli*. *Rna-a Publication of the Rna Society* **14**, 2223-2233.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. *Genetics* **131**, 479-491.

References

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7**, 574-578.

Felsenstein J (1985) Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution* **39**, 783-791.

Felsenstein J (2007) PHYLIP (Phylogeny Interference Package) Department of Genetics, University of Washington, Seattle.

Filipowicz W, Pogacic V (2002) Biogenesis of small nucleolar ribonucleoproteins. *Current Opinion in Cell Biology* **14**, 319-327.

Fu YX, Chakraborty R (1998) Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics* **150**, 487-497.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular Ecology* **6**, 861-868.

Ganot P, Bortolin ML, Kiss T (1997a) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell* **89**, 799-809.

Ganot P, CaizerguesFerrer M, Kiss T (1997b) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Development* **11**, 941-956.

References

Gaspin C, Cavaille J, Erauso G, Bachellerie JP (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: Lessons from the *Pyrococcus* genomes. *Journal of Molecular Biology* **297**, 895-906.

Gelperin D, Horton L, Beckman J, Hensold J, Lemmon SK (2001) Bms1p, a novel GTP-binding protein, and the related Tsr1p are required for distinct steps of 40S ribosome biogenesis in yeast. *Rna-a Publication of the Rna Society* **7**, 1268-1283.

Ghosh T, Peterson B, Tomasevic N, Peculis BA (2004) *Xenopus* u8 snoRNA binding protein is a conserved nuclear decapping enzyme. *Molecular Cell* **13**, 817-828.

Gulyas G, Sramko G, Molnar A, *et al.* (2005) Nuclear ribosomal DNA ITS paralogs as evidence of recent interspecific hybridization in the genus *Ophrys* (Orchidaceae). *Acta Biologica Cracoviensia Series Botanica* **47**, 61-67.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic acids symposium series* **41**, 95-98.

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**.

Hancock JM, Vogler AP (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Molecluar Phylogenetics and Evolution* **14**, 366-374.

Harris SA (2002) Introduction of Oxford ragwort, *Senecio squalidus* L. (Asteraceae), to the United Kingdom. *Watsonia* **24**, 31-43.

References

Harris SA, Ingram R (1992a) Molecular systematics of the genus *Senecio* L. 1. Hybridization in a British polyploid complex. *Heredity* **69**, 1-10.

Harris SA, Ingram R (1992b) Molecular systematics of the genus *Senecio* L. 2. The origin of *S. vulgaris* L. *Heredity* **69**, 112-121.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B-Biological Sciences* **270**, 313-321.

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology* **54**.

Henras AK, Capeyrou R, Henry Y, Caizergues-Ferrer M (2004a) Cbf5p, the putative pseudouridine synthase of H/ACA-type snoRNPs, can form a complex with Gar1p and Nop10p in absence of Nhp2p and box H/ACA snoRNAs. *Rna-a Publication of the Rna Society* **10**, 1704-1712.

Henras AK, Dez C, Henry Y (2004b) RNA structure and function in C/D and H/ACA s(no)RNPs. *Current Opinion in Structural Biology* **14**, 335-343.

Henras AK, Soudet J, Gerus M, *et al.* (2008) The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cellular and Molecular Life Sciences* **65**, 2334-2359.

Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* **24**, 158-164.

Hochschartner G (2006) *Molecular biogeography of the Bulbophyllum lobbii Lindl. complex (sect. Sestochilos, Orchidaceae) in Southeast Asia* Master thesis, Paris-Lodron-University of Salzburg.

References

Hollingsworth ML, Clark AA, Forrest LL, *et al*. (2009a) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources* **9**, 439-457.

Hollingsworth PM, Ennos RA (2004) Neighbour joining trees, dominant markers and population genetic structure. *Heredity* **92**, 490-498.

Hollingsworth PM, Forrest LL, Spouge JL, *et al.* (2009b) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12794-12797.

Huang ZP, Chen CJ, Zhou H, Li BB, Qu LH (2007) A combined computational and experimental analysis of two families of snoRNA genes from *Caenorhabditis elegans*, revealing the expression and evolution pattern of snoRNAs in nematodes. *Genomics* **89**, 490-501.

Huang ZP, Zhou H, He HL, *et al.* (2005) Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. *Rna-a Publication of the Rna Society* **11**, 1303-1316.

Huang ZP, Zhou H, Liang D, Qu LH (2004) Different expression strategy: Multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. *Journal of Molecular Biology* **341**, 669-683.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322-1332.

References

Huff DR, Peakall R, Smouse PE (1993) RAPD variation within and among natural populations of outcrossing Buffalograss [*Buchloe dactyloides* (Nutt) Engelm]. *Theoretical and Applied Genetics* **86**, 927-934.

Hughes JMX (1996) Functional base-pairing interaction between highly conserved elements of U3 small nucleolar RNA and the small ribosomal subunit RNA. *Journal of Molecular Biology* **259**, 645-654.

Hyten DL, Cannon SB, Song Q, *et al.* (2010) High-throughput SNP discovery deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 1-8.

Ingram R, Weir J, Abbott RJ (1980) New evidence concerning the origin of inland radiate groundsel, *S vulgaris* L var *hibernicus* Syme. *New Phytologist* **84**, 543-546.

Irwin JA, Abbott RJ (1992) Morphometric and isozyme evidence for the hybrid origin of a new tetraploid radiate groundsel in York, England. *Heredity* **69**, 431-439.

James JK, Abbott RJ (2005) Recent, allopatric, homoploid hybrid speciation: The origin of *Senecio squalidus* (Asteraceae) in the British Isles from a hybrid zone on Mount Etna, Sicily. *Evolution* **59**, 2533-2547.

Jarmolowski A, Zagorski J, Li HV, Fournier MJ (1990) Identification of essential elements in U14 RNA of *Saccharomyces cerevisiae*. *Embo Journal* **9**, 4503-4509.

Kadereit JW (1984) Studies on the Biology of *Senecio vulgaris* L ssp *denticulatus* (of-Muell) Pd-Sell. *New Phytologist* **97**, 681-689.

Kadereit JW, Uribe-Convers S, Westberg E, Comes HP (2006) Reciprocal hybridization at different times between *Senecio flavus* and *Senecio glaucus* gave rise to two

polyploid species in north Africa and south-west Asia. *New Phytologist* **169**, 431-441.

Kane NC, Cronk Q (2008) Botany without borders: barcoding in focus. *Molecular Ecology* **17**, 5175-5176.

Karbstein K, Doudna JA (2006) GTP-dependent formation of a ribonucleoprotein subcomplex required for ribosome biogenesis. *Journal of Molecular Biology* **356**, 432-443.

Karbstein K, Jonas S, Doudna JA (2005) An essential GTPase promotes assembly of preribosomal RNA processing complexes. *Molecular Cell* **20**, 633-643.

Kelchner SA (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* **87**, 482-498.

Kim M, Cui ML, Cubas P, *et al.* (2008) Regulatory genes control a key morphological and ecological trait transferred between species. *Science* **322**, 1116-1119.

Kimmel M, Chakraborty R (1996) Measures of variation at DNA repeat loci under a general stepwise mutation model. *Theoretical Population Biology* **50**, 345-367.

Kishore S, Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311**, 230-232.

Kiss-Laszlo Z, Henry Y, Kiss T (1998) Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *Embo Journal* **17**, 797-807.

References

Kiss AM, Jady BE, Bertrand E, Kiss T (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Molecular and Cellular Biology* **24**, 5797-5807.

Kohn MH, Wayne RK (1997) Facts from feces revisited. *Trends in Ecology & Evolution* **12**, 223-227.

Koopman WJM (2005) Phylogenetic signal in AFLP data sets. *Systematic Biology* **54**, 197-217.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8369-8374.

Kruszka K, Barneche F, Guyot R, *et al.* (2003) Plant dicistronic tRNA-snoRNA genes: a new mode of expression of the small nucleolar RNAs processed by RNase Z. *Embo Journal* **22**, 621-632.

Kuhner MK, Felsenstein J (1994) Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11**, 459-468.

Lafontaine DLJ, Tollervey D (1998) Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends in Biochemical Sciences* **23**, 383-388.

Lafontaine DLJ, Tollervey D (2001) The function and synthesis of ribosomes. *Nature Reviews Molecular Cell Biology* **2**, 514-520.

Lahaye R, Van der Bank M, Bogarin D, *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2923-2928.

References

Le Clerc-Blain J, Starr JR, Bull RD, Saarela JM (2010) A regional approach to plant DNA barcoding provides high species resolution of sedges (*Carex* and *Kobresia*, Cyperaceae) in the Canadian Arctic Archipelago. *Molecular Ecology Resources* **10**, 69-91.

Le Roux JJ, Wieczorek AM, Ramadan MM, Tran CT (2006) Resolving the native provenance of invasive fireweed (*Senecio madagascariensis* Poir.) in the Hawaiian Islands as inferred from phylogenetic analysis. *Diversity and Distributions* **12**, 694-702.

Leader DJ, Clark GP, Watters J, *et al.* (1997) Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *Embo Journal* **16**, 5742-5751.

Leader DJ, Clark GP, Watters J, *et al.* (1999) Splicing-independent processing of plant box C/D and box H/ACA small nucleolar RNAs. *Plant Molecular Biology* **39**, 1091-1100.

Leppik M, Peil L, Kipper K, Liiv A, Remme J (2007) Substrate specificity of the pseudouridine synthase RluD in *Escherichia coli*. *Febs Journal* **274**, 5759-5766.

Li HV, Zagorski J, Fournier MJ (1990) Depletion of U14 small nuclear-RNA (Snr128) disrupts production of 18S ribosomal-RNA in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **10**, 1145-1152.

Li W, Jiang G, Zeng DB, Jin YX (2007) Identification of six new box C/D snoRNA gene clusters from rice. *Iubmb Life* **59**, 664-674.

Liang D, Zhou H, Zhang P, *et al.* (2002) A novel gene organization: intronic snoRNA gene clusters from *Oryza sativa*. *Nucleic Acids Research* **30**, 3262-3272.

References

Liang WQ, Fournier MJ (1995) U14 base-pairs with 18S ribosomal-RNA - a novel snoRNA interaction required for ribosomal-RNA processing. *Genes & Development* **9**, 2433-2443.

Liston A, Kadereit JW (1995) Chloroplast DNA evidence for introgression and long-distance dispersal in the desert annual *Senecio flavus* (Asteraceae). *Plant Systematics and Evolution* **197**, 33-41.

Liu MH, Busch RK, Buckley B, Reddy R (1992) Characterization of antibodies against methyl-pppn cap structure - plant U3 small nucleolar RNA is recognized by these antibodies. *Nucleic Acids Research* **20**, 4299-4304.

Lowe AJ, Abbott RJ (1996) Origins of the new allopolyploid species *Senecio cambrensis* (asteraceae) and its relationship to the Canary Islands endemic S*enecio teneriffae*. *American Journal of Botany* **83**, 1365-1372.

Lowe AJ, Abbott RJ (2000) Routes of origin of two recently evolved hybrid taxa: *Senecio vulgaris* var. *hibernicus* and York radiate groundsel (Asteraceae). *American Journal of Botany* **87**, 1159-1167.

Lowe AJ, Harris S, Ashton P (2004) *Ecological Genetics: Design, Analysis, and Application* Blackwell Publishing.

Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168-1171.

Mai JC, Coleman AW (1997) The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *Journal of Molecular Evolution* **44**, 258-271.

References

Makarova JA, Kramerov DA (2007) Small nucleolar RNA. *Molecular Biology* **41**, 214-226.

Marker C, Zemann A, Terhorst T, *et al.* (2002) Experimental RNomics: Identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Current Biology* **12**, 2002-2013.

Markmann M, Tautz D (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1917-1924.

Matsuzaki H, Loi H, Dong S, *et al.* (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Research* **14**, 414-425.

Matthee CA, Eick G, Willows-Munro S, *et al.* (2007) Indel evolution of mammalian introns and the utility of non-coding nuclear markers in eutherian phylogenetics. *Molecular Phylogenetics and Evolution* **42**, 827-837.

Maxwell ES, Fournier MJ (1995) The small nucleolar RNAs. *Annual Review of Biochemistry* **64**, 897-934.

McDade LA (1992) Hybrids and phylogenetic systematics. 1. The impact of hybrids on cladistic-analysis. *Evolution* **46**, 1329-1346.

McGregor CE, Lambert CA, Greyling MM, Louw JH, 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.

References

Meier UT (2005) The many facets of H/ACA ribonucleoproteins. *Chromosoma* **114**, 1-14.

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology* **75**, 35-45.

Merchant SS, Prochnik SE, Vallon O, *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-251.

Mereau A, Fournier R, Gregoire A, *et al.* (1997) An in vivo and in vitro structure-function analysis of the Saccharomyces cerevisiae U3A snoRNP: Protein-RNA contacts and base-pair interaction with the pre-ribosomal RNA. *Journal of Molecular Biology* **273**, 552-571.

Meudt H, Clarke A (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science* **12**, 106-117.

Michot B, Joseph N, Mazan S, Bachellerie JP (1999) Evolutionarily conserved structural features in the ITS2 of mammalian pre-rRNAs and potential interactions with the snoRNA U8 detected by comparative analysis of new mouse sequences. *Nucleic Acids Research* **27**, 2271-2282.

Miklos I, Lunter GA, Holmes I (2004) A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution* **21**, 529-540.

Milton JJ (2009) *Phylogenetic analyses and taxonomic studies of Senecioninae: Southern African Senecio section Senecio.* PhD thesis, St. Andrews University.

References

Monaghan MT, Balke M, Gregory TR, Vogler AP (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1925-1933.

Mort ME, Crawford DJ, Archibald JK, O'Leary TR, Santos-Guerra A (2010) Plant DNA barcoding: A test using Macaronesian taxa of *Tolpis* (Asteraceae). *Taxon* **59**, 581-587.

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics* **8**, 6-21.

Nahkuri S, Taft RJ, Korbie DJ, Mattick JS (2008) Molecular evolution of the HBII-52 snoRNA cluster. *Journal of Molecular Biology* **381**, 810-815.

Nordborg M, Hu TT, Ishino Y, *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *Plos Biology* **3**, 1289-1299.

Nordenstam B (2003) Recent progress in Senecioneae taxonomy. *Compositae Newsletter* **40**, 26.

Nordenstam B (2007) Tribe Senecioneae. In: *Flowering Plants. Eudictos. Asterales* (eds. Kadereit JW, Jeffrey C), pp. 208-241. Springer, Berlin.

Ofengand J, Malhotra A, Remme J, *et al.* (2001) Pseudouridines and psendouridine synthases of the ribosome. *Cold Spring Harbor Symposia on Quantitative Biology* **66**, 147-159.

Olson M, Hood L, Cantor C, Botstein D (1989) A common language for physical mapping of the human genome. *Science* **245**, 1434-1435.

References

Omer AD, Ziesche S, Decatur WA, Fournier MJ, Dennis PP (2003) RNA-modifying machines in archaea. *Molecular Microbiology* **48**, 617-629.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Peculis BA (1995) U8 and U14 snoRNAs are essential for pre-ribosomal-RNA processing in Vertebrates. *Molecular Biology of the Cell* **6**, 1130-1130.

Peculis BA, DeGregorio S, McDowell K (2001) The U8 snoRNA gene family: identification and characterization of distinct, functional U8 genes in *Xenopus*. *Gene* **274**, 83-92.

Peculis BA, Steitz JA (1993) Disruption of U8 nucleolar snRNA inhibits 5.8s and 28s ribosomal-RNA processing in the *Xenopus*-oocyte. *Cell* **73**, 1233-1245.

Pelser PB, Nordenstam B, Kadereit JW, Watson LE (2007) An ITS phylogeny of tribe Senecioneae (Asteraceae) and a new delimitation of *Senecio* L. *Taxon* **56**, 1077-1104.

Pennisi E (2007) Taxonomy. Wanted: A barcode for plants. *Science* **318**, 190-191.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847-859.

Pons J, Vogler A (2006) Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. *Cladistics* **22**, 144-156.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

References

Qu LH, Henras A, Lu YJ, *et al.* (1999) Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Rat1p and RNase III in yeast. *Molecular and Cellular Biology* **19**, 1144-1158.

Qu LH, Meng Q, Zhou H, Chen YQ (2001) Identification of 10 novel snoRNA gene clusters from *Arabidopsis thaliana*. *Nucleic Acids Research* **29**, 1623-1630.

Radford IJ, Muller P, Fiffer S, Michael PW (2000) Genetic relationships between Australian Fireweed and South African and Madagascan populations of *Senecio madagascariensis* Poir. and closely related *Senecio* species. *Australian Systematic Botany*, 409-423.

Reddy R, Singh R, Shimba S (1992) Methylated cap structures in eukaryotic RNAs - structure, synthesis and functions. *Pharmacology & Therapeutics* **54**, 249-267.

Reichow SL, Hamma T, Ferre-D'Amare AR, Varani G (2007) The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Research* **35**, 1452-1464.

Riedel N, Wise JA, Swerdlow H, Mak A, Guthrie C (1986) Small nuclear RNAs from *Saccharomyces cerevisiae* - unexpected diversity in abundance, size, and molecular complexity. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 8097-8101.

Rieseberg LH, Beckstromsternberg S, Doan K (1990) *Helianthus annuus* ssp *texanus* has chloroplast DNA and nuclear ribosomal-RNA genes of *Helianthus debilis* ssp *cucumerifolius*. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 593-597.

Rieseberg LH, Soltis DE (1991) Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5**, 65-84.

References

Rotmistrovsky K, Jang W, Schuler GD (2004) A web server for performing electronic PCR. *Nucleic Acids Research* **32**, W108-W112.

Rozhdestvensky TS, Tang TH, Tchirkova IV, *et al.* (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Research* **31**, 869-877.

Russell AG, Schnare MN, Gray MW (2006) A large collection of compact box C/D snoRNAs and their isoforms in *Euglena gracilis*: Structural functional and evolutionary insights. *Journal of Molecular Biology* **357**, 1548-1565.

Saitou N, Nei M (1987) The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425.

Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide-sequences of primates. *Molecular Biology and Evolution* **11**, 504-512.

Samarsky DA, Fournier MJ, Singer RH, Bertrand E (1998) The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. *Embo Journal* **17**, 3747-3757.

SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 1460-1465.

Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1879-1888.

References

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1805-1811.

Schattner P, Decatur WA, Davis CA, *et al.* (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Research* **32**, 4281-4296.

Schlotterer C (2004) The evolution of molecular markers - just a matter of fashion? *Nature Reviews Genetics* **5**, 63-69.

Schmitz J, Zemann A, Churakov G, *et al.* (2008) Retroposed SNOfall - A mammalian-wide comparison of platypus snoRNAs. *Genome Research* **18**, 1005-1010.

Schuler GD (1997) Sequence mapping by electronic PCR. *Genome Research* **7**, 541-550.

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany* **94**, 275-288.

Shimba S, Buckley B, Reddy R, Kiss T, Filipowicz W (1992) Cap structure of U3 small nucleolar RNA in animal and plant-cells is different - gamma-monomethyl phosphate cap Structure in plant RNA. *Journal of Biological Chemistry* **267**, 13772-13777.

Shneyer VS (2009) DNA barcoding is a new approach in comparative genomics of plants. *Russian Journal of Genetics* **45**, 1267-1278.

References

Simmons M, Müller K, Norton A (2007) The relative performance of indel-coding methods in simulations. *Molecular Phylogenetics and Evolution* **44**, 724-740.

Singh R, Reddy R (1989) Gamma-monomethyl phosphate - a cap structure in spliceosomal U6 small nuclear RNA. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 8280-8283.

Small RL, Cronn RC, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* **17**, 145-170.

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: Choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* **85**, 1301-1315.

Soltis DE, Soltis PS, Chase MW, *et al.* (2000) Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Botanical Journal of the Linnean Society* **133**, 381-461.

Soltis PS, Plunkett GM, Novak SJ, Soltis DE (1995) Genetic variation in *Tragopogon* species - additional origins of the allotetraploids *T. mirus* and *T. miscellus* (Compositae). *American Journal of Botany* **82**, 1329-1341.

Suh YB, Thien LB, Reeve HE, Zimmer EA (1993) Molecular evolution and phylogenetic implications of internal transcribed spacer sequences of ribosomal DNA in Winteraceae. *American Journal of Botany* **80**, 1042-1055.

Taberlet P, Griffin S, Goossens B, *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

References

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11030-11035.

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. 1. Distantly related species. *Journal of Molecular Evolution* **18**, 387-404.

Terns MP, Grimm C, Lund E, Dahlberg JE (1995) A common maturation pathway for small nucleolar RNAs. *Embo Journal* **14**, 4860-4871.

Terns MP, Terns RM (2002) Small nucleolar RNAs: Versatile trans-acting molecules of ancient evolutionary origin. *Gene Expression* **10**, 17-39.

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

Thompson JR, Zagorski J, Woolford JL, Fournier MJ (1988) Sequence and genetic-analysis of a dispensable 189 nucleotide snRNA from *Saccharomyces cerevisiae*. *Nucleic Acids Research* **16**, 5587-5601.

Thongjuea S, Ruanjaichon V, Bruskiewich R, Vanavichit A (2009) RiceGeneThresher: a web-based application for mining genes underlying QTL in rice genome. *Nucleic Acids Research* **37**, 996-1000.

References

Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum-likelihood alignment of DNA-sequences. *Journal of Molecular Evolution* **33**, 114-124.

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality - an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**, 3-16.

Tollervey D (1996) Small nucleolar RNAs guide ribosomal RNA methylation. *Science* **273**, 1056-1057.

Tomasevic N, Peculis B (1999) Identification of a U8 snoRNA-specific binding protein. *Journal of Biological Chemistry* **274**, 35914-35920.

Tomasevic N, Peculis BA (2002) *Xenopus* LSm proteins bind U8 snoRNA via an internal evolutionarily conserved octamer sequence. *Molecular and Cellular Biology* **22**, 4101-4112.

Torchet C, Badis G, Devaux F, *et al.* (2005) The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *Rna-a Publication of the Rna Society* **11**, 928-938.

Trow AH (1912) On the inheritance of certaun characters in the common groundsel - *Senecio vulgaris* - and its segregates. *Journal of Genetics* **2**, 239-276.

Tycowski KT, Shu M-D, Steitz JA (1994) Requirement for intron-encoded U22 small nucleolar RNA 18S ribosomal RNA maturation. *Science* **266**, 1558-1561.

Tycowski KT, You ZH, Graham PJ, Steitz JA (1998) Modification of U6 spliceosomal RNA is guided by other small RNAs. *Molecular Cell* **2**, 629-638.

References

Venema J, Tollervey D (1999) Ribosome synthesis in *Saccharomyces cerevisiae*. *Annual Review of Genetics* **33**, 261-311.

Watkins NJ, Dickmanns A, Luhrmann R (2002) Conserved stem II of the box C/D motif is essential for nucleolar localization and is required, along with the 15.5K protein, for the hierarchical assembly of the box C/D snoRNP. *Molecular and Cellular Biology* **22**, 8342-8352.

Watkins NJ, Lemm I, Ingelfinger D, *et al.* (2004) Assembly and maturation of the U3 snoRNP in the nucleoplasm in a large dynamic multiprotein complex. *Molecular Cell* **16**, 789-798.

Wegierski T, Billy E, Nasr F, Filipowicz W (2001) Bms1p, a G-domain-containing protein, associates with Rcl1p and is required for 18S rRNA biogenesis in yeast. *Rna-a Publication of the Rna Society* **7**, 1254-1267.

Wendel JF (2000) Genome evolution in polyploids. *Plant Molecular Biology* **42**, 225-249.

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. In: *Advances in Agronomy, Vol 78*, pp. 139-186.

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences of the United States of America* **92**, 280-284.

Xiao L-Q, Möller M, Zhu H (2010) High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: Incomplete concerted evolution and the origin of pseudogenes. *Molecular Phylogenetics and Evolution* **55**, 168-177.

Yamane K, Yano K, Kawahara T (2006) Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. *DNA Research* **13**, 197-204.

Yang JH, Zhang XC, Huang ZP, *et al.* (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Research* **34**, 5112-5123.

Yonemaru J-I, Ando T, Mizubayashi T, *et al.* (2009) Development of genome-wide simple sequence repeat markers using whole-genome shotgun sequences of *Sorghum* (*Sorghum bicolor* (L.) Moench). *DNA Research* **16**, 187-193.

You FM, Wanjugi H, Huo N, *et al.* (2010) RJPrimers: unique transposable element insertion junction discovery and PCR primer design for marker development. *Nucleic Acids Research* **38**, 313-320.

Young N, Healy J (2003) GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics* **4**, 1-6.

Yu J, Hu SN, Wang J, *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* **296**, 79-92.

Zemann A, op de Bekke A, Kiefmann M, Brosius J, Schmitz J (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Research* **34**, 2676-2685.

Appendix

Chapter 1

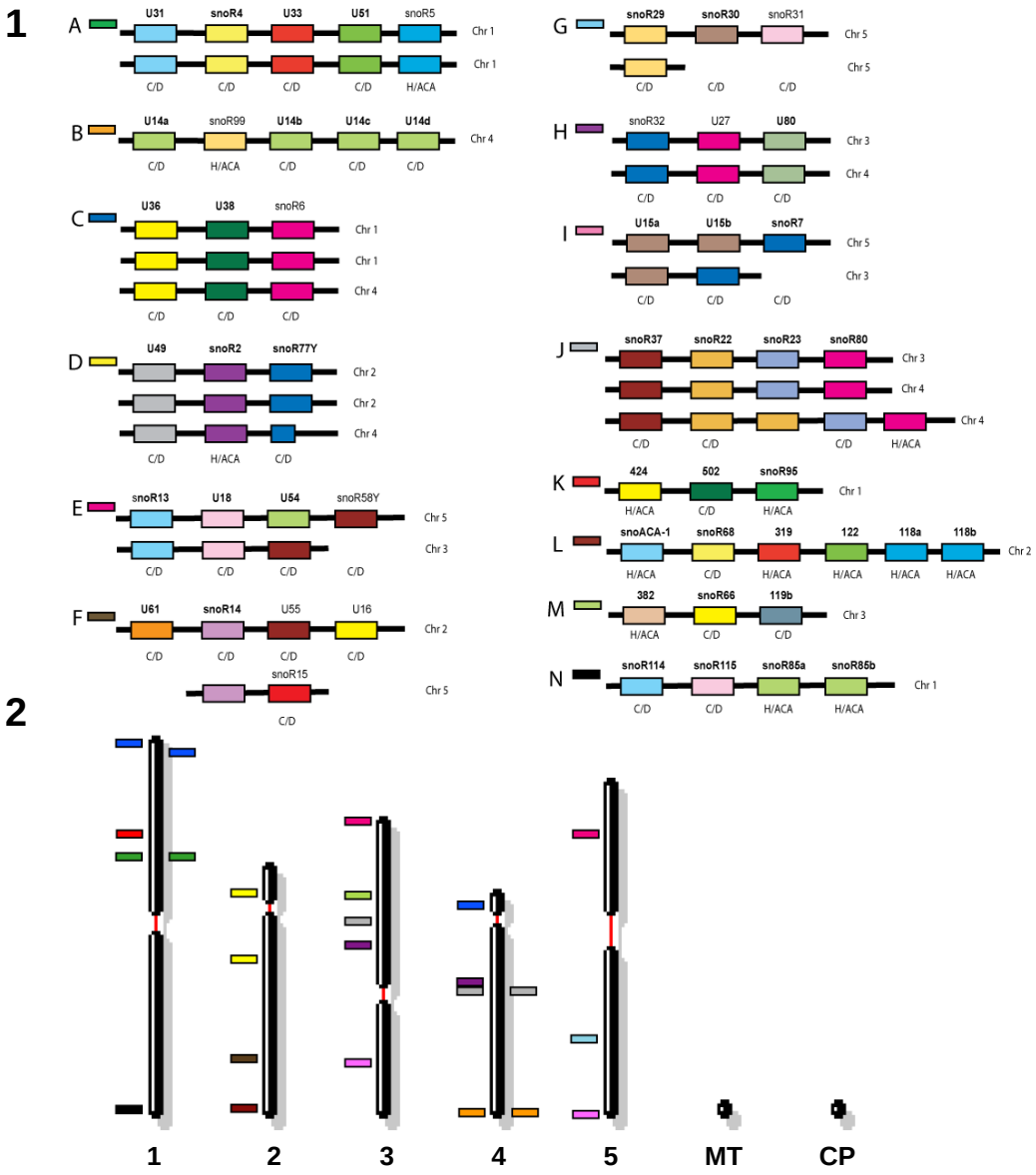


Figure A.1: Investigated snoRNA genes and gene clusters (1) and their approximate location within the *Arabidopsis thaliana* genome (2). (A-J) Gene clusters characterized by Brown et al. (2001). (K-N) Single copy gene clusters recently discovered by Brown et al. (unpublished data). Genes are displayed by boxes of different colours. Genes which names are written in bold were used in this study. C/D and H/ACA below the genes indicate the snoRNA gene type. The coloured boxes next to the cluster letters (A-M) in

(1) identify their cluster location(s) within the genome on the chromosome map (2). Chromosome map and positions of the gene cluster were obtained by using ePCR and map viewer (http://www.ncbi.nlm.nih.gov).

Table A.1: ITS sequences used for the reconstruction of the evolutionary relationships between Senecio species.

Species	Sequence ID	Species	Sequence ID
S. aethnensis	gi|7636057	*S. mohavensis* subsp. *breviflorus*	gi|18642591
S. chrysanthemifolius	gi|7636058	*S. mohavensis* subsp. *mohavensis*	gi|18642592
S. flavus	gi|18642572	*S. squalidus*	gi|21425756
S. flavus subsp. *flavus*	gi|7636060	*S. squalidus* subsp. *araneosus*	gi|7636082
S. glaucus subsp. *glaucus*	gi|18642596	*S. squalidus* subsp. *squalidus*	gi|7636081
S. glaucus supsp. *coronopifolius*	gi|18642595	*S. vulgaris*	gi|156754245
S. inaequidens	gi|84043214	*S. vulgaris* subsp. *denticulatus 1*	gi|7636219
S. madagascariensis 1	gi|84043205	*S. vulgaris* subsp. *denticulatus 2*	gi|7636220
S. madagascariensis 2	gi|84043206	*S. vulgaris* subsp. *vulgaris*	gi|7636218

Chapter 3

Table A.2 Sequences used for alignments and gene organization reconstruction.

Species	Sequence ID	Source
Cluster A		
424		
Arabidopsis thaliana	gi:12321165	genomic
Cichorium intybus	gi:124575509	EST
Citrus aurantiifolia	gi:188367092/ gi:188254127	EST
Citrus sinensis	gi:56585714	EST
Medicago truncatula	gi:209863214	genomic
Vitis vinifera	gi:147817707	genomic
502		
Arabidopsis thaliana	gi:12321165	genomic
Citrus sinensis	gi:56585714	EST
Citrus aurantiifolia	gi:188367092	EST
Cleome spinosa	gi:255773577	EST
snoR95		
Arabidopsis thaliana	gi:12321165	genomic
Citrus sinensis	gi:56585714	EST
Citrus aurantiifolia	gi:188367092	EST
complete cluster sequence		
Arabidopsis thaliana	gi:12321165	genomic
Citrus sinensis	gi:56585714	EST
Citrus aurantiifolia	gi:188367092	EST
Cleome spinosa	gi:255773577	EST
Cluster B		
snoACA-1		
Arabidopsis thaliana	gi:240254678	genomic
Brassica rapa subsp. *pekinensis*	gi:199580275	genomic
snoR68		
Arabidopsis thaliana	gi:240254678	genomic
Brassica oleracea	gi:26784985	genomic
Brassica rapa subsp. *pekinensis*	gi:199580275	genomic
Elaeis guineensis	gi:56930801	EST
Euphorbia esula	gi:76858378	EST
Medicago truncatula	gi:83665967	EST
Oryza sativa	gi:33380410/ gi:33380495	genomic
Populus tremula x Populus tremuloides	gi:24080304	EST
Solanum lycopersicum	gi:4381187	EST
Solanum tuberosum	gi:53699676	EST
319		
Arabidopsis thaliana	gi:240254678	genomic
Brassica rapa subsp. *pekinensis*	gi:199580275	genomic
Acorus americanus	gi:74103691	EST
Glycine max	gi:18731564	EST
Lactuca virosa	gi:84017452	EST
Medicago truncatula	gi:83665967	EST
Raphanus sativus	gi:166140359	EST

122		
Arabidopsis thaliana	gi:240254678	genomic
Brassica rapa subsp. pekinensis	gi:199580275	genomic
Citrus unshiu	gi:209935776	EST
Glycine max	gi:192301296	EST
Guizotia abyssinica	gi:211722687	EST
Populus nigra	gi:161935253	EST
Populus tremula x Populus tremuloides	gi:60698243	EST
Raphanus sativus	gi:166140359	EST
118a/b		
Arabidopsis thaliana	gi:240254678	genomic
Raphanus sativus	gi:166143527	EST
Brassica rapa subsp. *pekinensis*	gi:156808274	genomic
Carica papaya	gi:186804059	EST
Citrullus lanatus	gi:198410797	EST
Lactuca saligna	gi:83803504	EST
Mimulus guttatus	gi:53844705	EST
Phaseolus coccineus	gi:27402212	EST
Lactuca virosa	gi:84029984	EST
Complete cluster sequence		
Arabidopsis thaliana	gi:240254678	genomic
Brassica rapa subsp. *pekinensis*	gi:199580275	genomic
Raphanus sativus	gi:166140359	EST
Cluster C		
382		
Arabidopsis thaliana	gi:240255695	genomic
Brassica rapa subsp. pekinensis	gi:199580032	genomic
Raphanus sativus	gi:166149116	EST
Raphanus raphanistrum subsp. *landra*	gi:166125710	EST
snoR66		
Arabidopsis thaliana	gi:240255695	genomic
Brassica oleracea genomic	gi:23523856	genomic
Brassica rapa subsp. *pekinensis*	gi:57900807	genomic
Glycine max	gi:26044093	EST
Lactuca perennis	gi:83886679	EST
Lactuca sativa	gi:83992923	EST
Lactuca serriola	gi:22438909	EST
Lactuca virosa	gi:84009164	EST
Medicago truncatula	gi:13780193/ gi:86361386	EST/genomic
Oryza sativa	gi:27548534	EST
Oryza sativa a/b/c/d/g/h	gi:27527582/ gi:27527583/ gi:27527584/ gi:27527585/ gi:27527586/ gi:27527587	genomic
Picea glauca	gi:49141414	EST
Populus tremula x Populus tremuloides	gi:24076600	EST
Populus trichocarpa	gi:52386830	EST
Pseudotsuga menziesii var. *menziesii*	gi:47146297	EST
Salvia miltiorrhiza	gi:51958830	EST
Solanum lycopersicum	gi:62927503	EST

Solanum tuberosum	gi:10448481/ gi:60696644/ gi:60706804/ gi:21915632/ gi:20170484	EST
Vitis vinifera	gi:110369092/ gi:110390903/ gi:83276605	EST
Zea mays	gi:50331559	EST
119b		
Arabidopsis thaliana	gi:240255695	genomic
Brassica napus	gi:65285354	EST
Brassica rapa subsp. *pekinensis*	gi:37621417/ gi:57900807	EST/genomic
Glycine max	gi:26044093	EST
Gossypium arboreum	gi:21092412	EST
Helianthus tuberosus	gi:125445895	EST
Lactuca perennis	gi:83886679	EST
Lactuca sativa	gi:83992923	EST
Lactuca serriola	gi:22438909	EST
Lactuca virosa	gi:84009164	EST
Medicago truncatula	gi:13780193/ gi:86361386	EST/genomic
Populus tremula	gi:60696644/ gi:60706804	EST
Populus tremula x Populus tremuloides	gi:24076600	EST
Populus trichocarpa	gi:52386830	EST
Raphanus raphanistrum subsp. *landra*	gi:166125710	EST
Raphanus sativus	gi:166145717	EST
Solanum lycopersicum	gi:62927503	EST
Vitis vinifera	gi:110390903/ gi:83277108/ gi:83276605	EST
Complete cluster sequence		
Arabidopsis thaliana	gi:240255695	genomic
Brassica rapa subsp. *pekinensis*	gi:199580032	genomic
Cyamopsis tetragonoloba	gi:117894741	EST
Guizotia abyssinica	gi:211705865	EST
Helianthus tuberosus	gi:125445895	EST
Lactuca sativa	gi:90521404	EST
Lactuca serriola	gi:22438909	EST
Lactuca virosa	gi:84009164	EST
Medicago truncatula	gi:152924799/ gi:86361386/ gi:13780193	genomic
Oryza sativa	gi:54291824/ gi:28564732	genomic
Populus petioles	gi:60706804	EST
Raphanus raphanistrum subsp. *landra*	gi:166125710	EST
Raphanus sativus	gi:166145717	EST
Vigna unguiculata	gi:190455218	EST
Vitis vinifera	gi:110390903/ gi:83276605	EST
Populus tremula	gi:60696644/ gi:60706804	EST
Populus trichocarpa	gi:52386830	EST
Cluster D		
snoR37		
Arabidopsis thaliana	gi:240255695/ gi:240256243	genomic
Brassica napus	gi:73674964/ gi:150928304	genomic
Brassica oleracea	gi:17748447	genomic
Brassica rapa subsp. *pekinensis*	gi:110797191	genomic
Carica papaya	gi:186764559	EST

Euphorbia esula	gi:76853977	EST
Glycine max	gi:15203746/ gi:31309888/ gi:33390105	EST
Gossypium hirsutum	gi:109869959	EST
Ipomoea nil	gi:74383573	EST
Lactuca sativa	gi:22234984	EST
Lactuca serriola	gi:83917712	EST
Lactuca virosa	gi:84010025	EST
Medicago truncatula	gi:7562914	EST
Nicotiana tabacum	gi:156666631	EST
Physcomitrella patens subsp. *patens*	gi:18361319	EST
Populus alba x Populus tremula	gi:57890243	EST
Raphanus raphanistrum subsp. *maritimus*	gi:166100767	EST
Raphanus sativus	gi:167451966	EST
Solanum lycopersicum	gi:14684154	EST
Solanum tuberosum	gi:53776780	EST
Vitis vinifera	gi:110429485/ gi:110721679	EST
snoR80		
Arabidopsis thaliana	gi:240255695/ gi:240256243	genomic
Beta vulgaris	gi:21333682	EST
Brassica napus	gi:189101690	EST
Brassica rapa subsp. pekinensis	gi:110797191	genomic
Carthamus tinctorius	gi:125382826	EST
Cistus creticus subsp. *creticus*	gi:182408457	EST
Cucumis melo subsp. *melo*	gi:157723088	EST
Gossypium hirsutum	gi:164324599	EST
Lactuca sativa	gi:90507839	EST
Lactuca serriola	gi:83908442	EST
Lotus japonicus	gi:29122726	genomic
Medicago truncatula	gi:209567374/ gi:144225814	EST/genomic
Panicum virgatum	gi:198296334	EST
Phaseolus vulgaris	gi:171544740	EST
Phyllostachys edulis	gi:242375504	EST
Raphanus sativus	gi:156162231	EST
Raphanus sativus var. *oleiformis*	gi:166134883	EST
Solanum tuberosum	gi:21371550	EST
Vitis vinifera	gi:30305307/ gi:123663939	EST/genomic
Zea mays	gi:32944143	EST
Complete cluster sequence		
Arabidopsis thaliana	gi:240255695/ gi:240256243	genomic
Barnadesia spinosa	gi:211666183	EST
Beta vulgaris	gi:21333682	EST
Brassica napus	gi:150076027	EST
Brassica rapa subsp. *pekinensis*	gi:110797191/ gi:150155151	genomic
Carica papaya	gi:186764559	EST
Carthamus tinctorius	gi:125382826	EST
Centaurea maculosa	gi:124654980	EST
Citrus sinensis	gi:188232298	EST
Euphorbia esula	gi:76853977	EST
Glycine max	gi:58021921	EST

Gossypium hirsutum	gi:109869959	EST
Lactuca sativa	gi:22234984/ gi:90503277	EST
Lotus japonicus	gi:29122726	genomic
Medicago truncatula	gi:189458711/ gi:144225814	genomic
Oryza sativa	gi:49388339	genomic
Phaseolus vulgaris	gi:171544740	EST
Phyllostachys edulis	gi:242375504	EST
Populus alba x Populus tremula	gi:57890243	EST
Populus trichocarpa	gi:158749687	genomic
Raphanus raphanistrum subsp. *maritimus*	gi:166100767	EST
Raphanus sativus	gi:167450881	EST
Solanum lycopersicum	gi:182887681	genomic
Theobroma cacao	gi:212131282	EST
Tropaeolum majus	gi:215785731	EST
Vitis vinifera	gi:123663939	genomic
Cluster E		
snoR114		
Arabidopsis thaliana	gi:240254421	genomic
Brassica rapa subsp. *pekinensis*	gi:199579994	genomic
Citrus sinensis	gi:56586109	EST
Euphorbia esula	gi:76858228	EST
Fragaria vesca	gi:89548839	EST
Gossypium hirsutum	gi:164288770	
Helianthus annuus	gi:90442016	EST
Lactuca sativa	gi:90510312	EST
Lactuca serriola	gi:83910223/ gi:83921121	EST
Limonium bicolor	gi:56906983	EST
Lotus japonicus genomic	gi:185115103	genomic
Medicago truncatula	gi:13366608/ gi:60543399	EST/genomic
Mimulus guttatus	gi:238361331	EST
Oryza sativa	gi:42409361	genomic
Populus trichocarpa	gi:24069242/ gi:38598467	EST
Raphanus raphanistrum subsp. *landra*	gi:166125356	EST
Raphanus sativus	gi:167443322	EST
Ricinus communis	gi:111157100	EST
Saccharum officinarum	gi:35249149	EST
Triticum aestivum	gi:25193829	EST
Vitis vinifera	gi:71870793/ gi:147866745	EST/genomic
snoR15		
Arabidopsis thaliana	gi:240254421	genomic
Brassica rapa subsp. *pekinensis*	gi:199579994	genomic
Euphorbia esula	gi:76858228	EST
Festuca arundinacea	gi:74433438	EST
Fragaria vesca	gi:89548839	EST
Gossypium hirsutum	gi:109878620	EST
Lotus japonicus	gi:223434137	EST
Malus x domestica	gi:48110246/ gi:48113257	EST
Pinus taeda	gi:10681647	EST
Populus trichocarpa	gi:24069242/ gi:38598467	EST
Raphanus raphanistrum subsp. *landra*	gi:166125356	EST

Raphanus sativus	gi:156172344	EST
Vitis vinifera	gi:71870793/ gi:147866745	EST/genomic
snoR85a/b		
Arabidopsis thaliana	gi:240254421	genomic
Arachis hypogaea	gi:225615191	EST
Brassica rapa subsp. *pekinensis*	gi:199579994	genomic
Citrullus lanatus	gi:198410141	EST
Cucumis melo subsp. *agrestis*	gi:157707215	EST
Cyamopsis tetragonoloba	gi:117903472	EST
Euphorbia tirucalli	gi:58205853	EST
Festuca arundinacea	gi:74458026	EST
Ipomoea nil	gi:74417076	EST
Lotus japonicus	gi:29122723	genomic
Panicum virgatum	gi:197953108	EST
Raphanus raphanistrum subsp. *landra*	gi:166125356	EST
Raphanus sativus	gi:156172344	EST
Senecio vulgaris subsp. *vulgaris*	gi:89509231/ gi:89502815	EST
Solanum lycopersicum	gi:182887681	genomic
Solanum pennellii	gi:12636150	EST
Solanum tuberosum	gi:45290292	EST
Triphysaria versicolor	gi:159062187	EST
Zea mays	gi:50328560	EST
Complete cluster sequence		
Arabidopsis thaliana	gi:240254421	genomic
Brassica rapa subsp. *pekinensis*	gi:199579994	genomic
Euphorbia esula	gi:76858228	EST
Euphorbia tirucalli	gi:58205853	EST
Ipomoea nil	gi:74417076	EST
Lotus japonicus	gi:185115103	genomic
Raphanus raphanistrum subsp. *landra*	gi:166125356	EST
Raphanus sativus	gi:156172344	EST
Solanum pennellii	gi:12636150	EST
Triphysaria pusilla	gi:159666042	EST
Vitis vinifera	gi:123705899	genomic

Chapter 4

Analyses of fragment frequencies

Table A.3 to Table A.20: Fragment frequencies tables of the primer pairs examined. Only fragments with a within species frequency of at least 0.33 (moderate frequency fragments (mffs) are shown. Within species frequencies above 0.5 (high frequency fragments (hffs)) are shaded in grey.

Table A.3: Fragment frequencies of the U31/U51 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

		U31/U51											
Species	**N**	**163**	**182**	**192**	**207**	**284**	**311**	**340**	**444**	**480**	**500**	**516**	**628**
S. aethnensis	**12**			0.25		0.42		0.92		0.92			0.17
S. chrysanthemifolius	**12**			0.33		0.42		1.00		0.92			0.33
S. squalidus	**25**			0.48	0.04	0.36		0.96		0.88		0.08	0.16
S. vulgaris	**11**						0.91	1.00	0.73	1.00			
S. cambrensis	**12**						1.00	1.00	0.75	0.92			
S. madagascariensis	**8**	0.50	0.63		0.63	0.25	0.25	0.75		0.88	0.63	0.50	0.13

Table A.4: Fragment frequencies of the U14-3/U14-4 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

		U14-3/U14-4										
Species	**N**	**123**	**124**	**129**	**130**	**136**	**138**	**143**	**354**	**421**	**680**	**694**
S. aethnensis	10			0.90	0.70						0.90	0.90
S. chrysanthemifolius	14			0.93	0.36						0.93	1
S. squalidus	28			0.86	0.57						0.64	0.75
S. vulgaris	13			1.00							0.69	0.62
S. cambrensis	11			1.00	0.91						0.55	0.64
S. madagascariensis	9					0.33	0.89	0.44				
S. flavus	2	1.00	1.00						0.50	0.50		

Table A.5: Fragment frequencies of the U49/snoR2 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

		U49/SR2										
Species	**N**	**116**	**118**	**120**	**123**	**127**	**131**	**134**	**149**	**155**	**216**	**235**
S. aethnensis	11	0.64	0.36	0.36	0.73	0.73	1.00	1.00	0.45	0.36	0.45	
S. chrysanthemifolius	10			0.50	0.60	0.20	1.00	1.00		0.20	1.00	0.50
S. squalidus	10						1.00	1.00			0.50	0.30
S. vulgaris	10	0.60			0.90		1.00	1.00		0.10		0.20
S. cambrensis	11	0.55			0.73		1.00	1.00			0.55	0.27

Table A.6: Fragment frequencies of the snoR2/snoR77Y primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	SR2/SR77Y	
		93	157
S. aethnensis	6	1.00	
S. chrysanthemifolius	11	1.00	
S. squalidus	10	1.00	0.10
S. vulgaris	5	0.80	0.60
S. cambrensis	7	0.86	0.86

Table A.7: Fragment frequencies of the U49/snoR77Y primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	U49/SR77Y													
		116	120	122	127	131	134	155	233	303	308	391	440	444	553
S. aethnensis	9	0.89	1.00	0.89	0.89	1.00	0.89	0.78	0.11	0.89	1.00	0.89			0.78
S. chrysanthemifolius	11	0.45	1.00	1.00	0.18	1.00	1.00	0.18	0.18	0.45	0.55	0.55			0.18
S. squalidus	6		1.00	1.00		1.00	1.00			0.17	0.17	0.33			
S. vulgaris	6		0.83	0.67		1.00	1.00		0.17						
S. cambrensis	6		1.00	1.00		1.00	1.00	0.17	0.67				0.50	0.50	

Table A.8: Fragment frequencies of the snoR13/U18 (SR13/U18) primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Samples	N	snoR13/U18							
		92	98	102	106	110	116	120	127
S. aethnensis	10	0.20	0.90		0.10	0.40		1.00	
S. chrysanthemifolius	11	0.73	0.64		0.09	0.18		1.00	
S. squalidus	27	0.85	0.78	0.04	0.11	0.22	0.04	0.81	
S. vulgaris	11	1.00	0.36	0.18	0.36	0.27	0.18	0.82	
S. cambrensis	11	1.00	0.91	0.09		0.09		1.00	
S. teneriffae	3	1.00	1.00			0.33		1.00	
S. madagascariensis	9	0.11	0.22	0.67	0.44	1.00	0.44	0.11	0.89
S. flavus	3			1.00	1.00	1.00	1.00		

Table A.9: Fragment frequencies of the U18/U54 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	U18/U54						
		98	143	160	165	185	188	320
S. aethnensis	10	1.00		0.50	1.00			
S. chrysanthemifolius	11	1.00		0.09	1.00			
S. squalidus	18	1.00		0.44	0.78			
S. vulgaris	7	1.00	0.86		0.29			0.57
S. cambrensis	10	1.00			0.90	0.10	0.10	0.50
S. madagascariensis	9	1.00		0.22		0.89	0.67	

Table A.10: Fragment frequencies of the snoR13/U54 (SR13/U54) primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR13/U54					
		97	255	368	376	379	645
S. aethnensis	5	1.00		0.40		0.80	
S. chrysanthemifolius	8	1.00		0.75	0.38	0.75	0.50
S. squalidus	3	1.00				0.33	
S. vulgaris	5	1.00	0.20			1.00	
S. cambrensis	3	1.00	1.00			1.00	

Table A.11: Fragment frequencies of the U61/snoR14 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Samples	N	U61/snoR14								
		114	116	120	123	129	132	137	222	293
S. aethnensis	11		1.00	0.73	0.64		0.64			
S. chrysanthemifolius	12	0.08	1.00	0.50	0.92					
S. squalidus	27		1.00	1.00	1.00		0.04	0.22	0.15	
S. vulgaris	12		0.08	1.00	1.00			0.75	0.42	0.08
S. cambrensis	12		0.33	1.00	1.00		0.27	0.87		
S. teneriffae	3		0.67	1.00	1.00		0.33	0.33		
S. madagascariensis	9	0.44	0.33	0.44	0.44	0.67				
S. flavus	3				1.00					1.00

Table A.12: Fragment frequencies of the U80-1/U80-2 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	U80-1/U80-2					
		128	**138**	**150**	**286**	**296**	**321**
S. aethnensis	10		1.00		0.60		0.20
S. chrysanthemifolius	11		1.00				
S. squalidus	15	0.20	1.00				
S. vulgaris	9	0.11	0.78	0.11		0.11	
S. cambrensis	10	0.10	1.00				

Table A.13: Fragment frequencies of the U15/snoR7 (U15/SR7) primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	U15/snoR7	
		92	**157**
S. aethnensis	10	1.00	1.00
S. chrysanthemifolius	10	1.00	1.00
S. squalidus	5	1.00	1.00
S. vulgaris	4	1.00	1.00
S. cambrensis	4	1.00	1.00

Table A.14: Fragment frequencies of the snoR37/snoR22 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR37/snoR22												
		105	**120**	**126**	**132**	**136**	**142**	**150**	**156**	**330**	**343**	**360**	**372**	**405**
S. aethnensis	10	0.20	0.30	0.20	0.40	0.80	0.80	1.00	0.40	0.10	0.50	0.20	0.10	0.40
S. chrysanthemifolius	11	0.73	0.64	0.82		0.27		1.00		0.82	0.18	0.36	0.55	
S. squalidus	15	0.80	0.33		0.07	0.47	0.27	0.93		0.47	0.07	0.27	0.07	
S. vulgaris	11				0.55	1.00		0.73	0.82					0.18
S. cambrensis	11	0.64		0.09	0.55	1.00		0.73	0.45					

Table A.15: Fragment frequencies of the sno22/snoR23 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR22/snoR23										
		197	**202**	**208**	**215**	**219**	**224**	**229**	**426**	**440**	**480**	**544**
S. aethnensis	12	0.33	1.00	0.42		0.50	1.00		0.33	0.25		
S. chrysanthemifolius	12	0.42	1.00			0.92	0.92	0.25	0.08	0.33		
S. squalidus	30	0.07	1.00	0.17		0.77	0.97	0.03		0.47		
S. vulgaris	13		1.00	0.69		0.38	0.08	0.77			0.62	0.15
S. cambrensis	12		1.00	0.75		0.75	0.75	1.00	0.17	0.67	1.00	0.67
S. teneriffae	3		1.00	0.33			1.00	0.33		0.33		0.33
S. madagascariensis	9		0.67	0.22	1.00	0.44	0.33					
S. flavus	3		1.00									

Table A.16: Fragment frequencies of the sno37/snoR23 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR37/snoR23											
		92	**203**	**310**	**333**	**343**	**348**	**355**	**365**	**530**	**548**	**560**	**592**
S. aethnensis	9	0.89			0.67		0.33	0.56		0.33	0.67	0.11	0.11
S. chrysanthemifolius	11	0.73	0.09	0.55	0.27			0.82		0.73	0.91	0.82	
S. squalidus	6	0.83	0.33				0.17	0.50		0.67	0.67	0.17	
S. vulgaris	6	1.00											
S. cambrensis	6				1.00	0.83		0.50	0.67	0.83	0.83		1.00

Table A.17: Fragment frequencies of the sno66/119bR1 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR66/119bR1						
		100	**105**	**108**	**125**	**210**	**354**	**481**
S. aethnensis	**5**	1.00	0.60	0.20	1.00	1.00		
S. chrysanthemifolius	**4**	0.75	0.25	1.00	0.75	0.75		
S. squalidus	**8**	1.00	0.13	1.00	0.38			
S. vulgaris	**5**		0.20	0.80	0.60		1.00	0.80
S. cambrensis	**6**	0.83		1.00	0.50		0.67	0.67

Table A.18: Fragment frequencies of the sno66/119bR2 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR66/119bR2							
		97	**146**	**160**	**165**	**168**	**174**	**176**	**257**
S. aethnensis	5			1.00	0.60	0.40			1.00
S. chrysanthemifolius	5			1.00	0.20				1.00
S. squalidus	8			1.00		0.88	0.25		0.25
S. vulgaris	7		0.43						0.57
S. cambrensis	9			1.00		0.78			0.44
S. madagascariensis	4	0.75					0.75	0.75	

Table A.19: Fragment frequencies of the snoR114/snoR85 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR114/snoR85										
		97	**123**	**128**	**257**	**300**	**328**	**385**	**404**	**438**	**725**	**770**
S. aethnensis	5	1.00	0.20	0.60	0.80	0.40	0.20					0.60
S. chrysanthemifolius	5	1.00	1.00		0.80					1.00		0.20
S. squalidus	6	1.00	1.00								0.67	0.50
S. vulgaris	4	1.00	0.50				0.75	0.75	1.00		1.00	1.00
S. cambrensis	4	1.00	0.75			0.50	0.50		1.00		0.75	0.75

Table A.20: Fragment frequencies of the sno115/snoR85 primer combination within *Senecio* species. High frequency fragments are shaded in grey.

Species	N	snoR115/snoR85														
		97	**148**	**223**	**232**	**258**	**336**	**354**	**359**	**411**	**536**	**587**	**610**	**621**	**739**	**770**
S. aethnensis	5	1.00		0.40	0.20	1.00		0.20	1.00	1.00		1.00	0.80	1.00	0.60	0.40
S. chrysanthemifolius	5	1.00		0.60	0.20	1.00		0.60	1.00	1.00		1.00	1.00	1.00	1.00	0.60
S. squalidus	6	1.00			0.67	1.00		0.83	1.00	1.00	0.33	1.00	1.00	1.00	1.00	
S. vulgaris	4	1.00	0.50		0.50		0.75		1.00	1.00		1.00	1.00			1.00
S. cambrensis	4	1.00			0.50				1.00	1.00		1.00	1.00	1.00	1.00	1.00

NJ and PCO analyses

Figure A.2 to Figure A.15: NJ trees and PCO plots of the primer pairs examined.

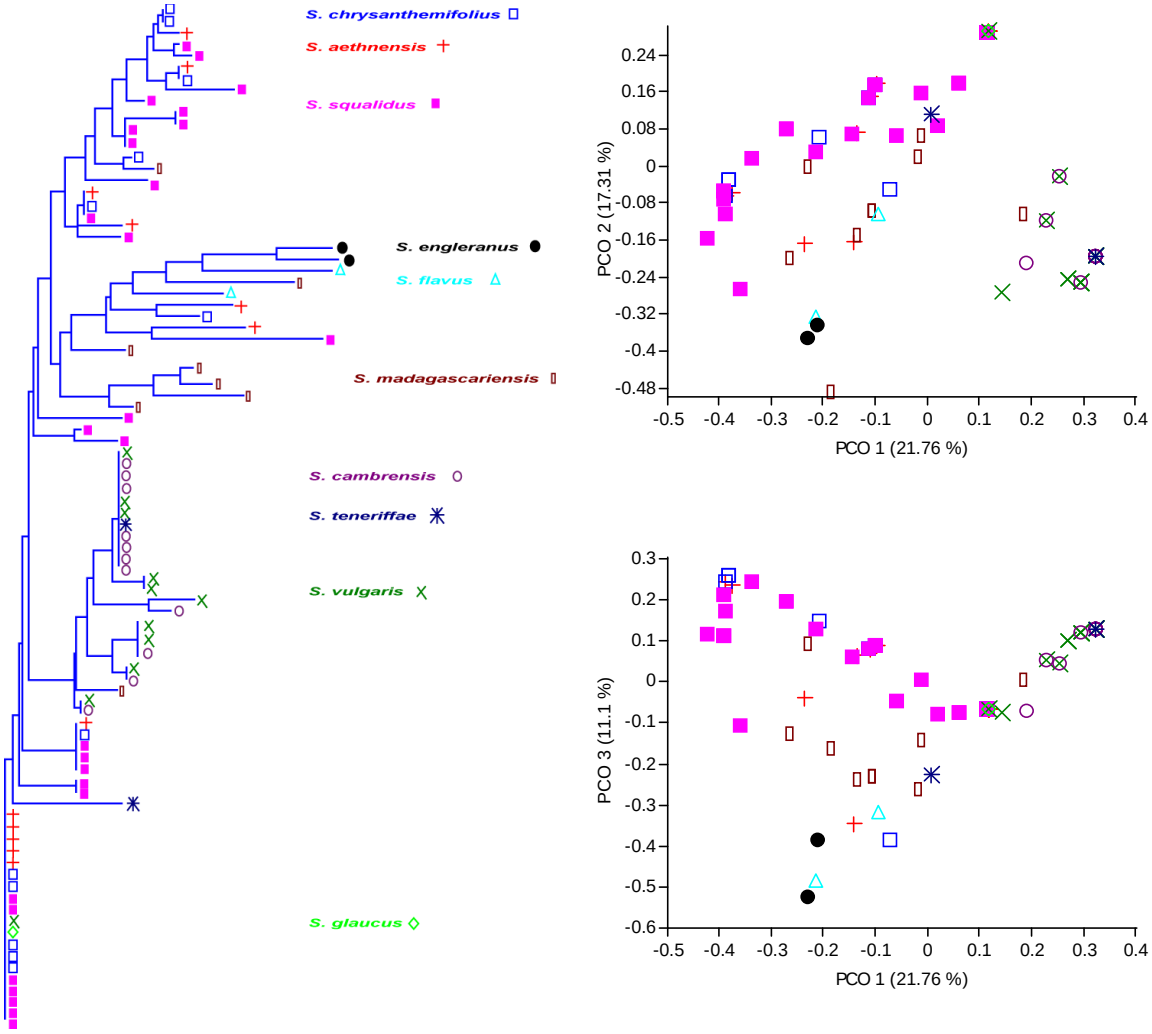


Figure A.2: NJ tree and PCO plots of *Senecio* sp. U31/U51 fragment profiles. NJ and PCO analyses of 85 samples made up of 10 species are based on fragment variation (31 fds) and dice genetic similarities of the U31F/U51R primer pair. The first three axis of the PCO explain 50.17 % of the variation within the dataset.

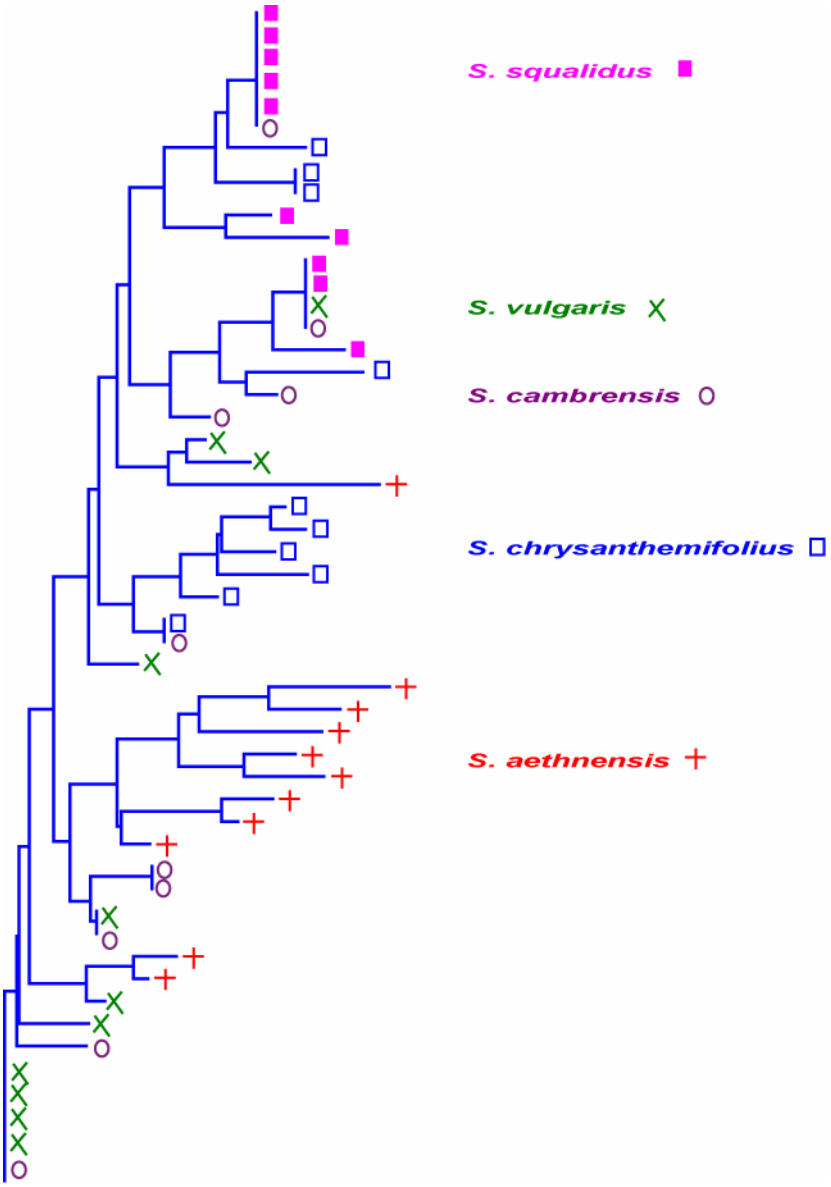


Figure A.3: NJ tree of *Senecio* sp. U49/snoR2 fragment profiles. NJ analysis of 52 samples made up of 6 species is based on fragment variation (24 fds) and dice genetic similarities of the U49F/snoR2R primer pair.

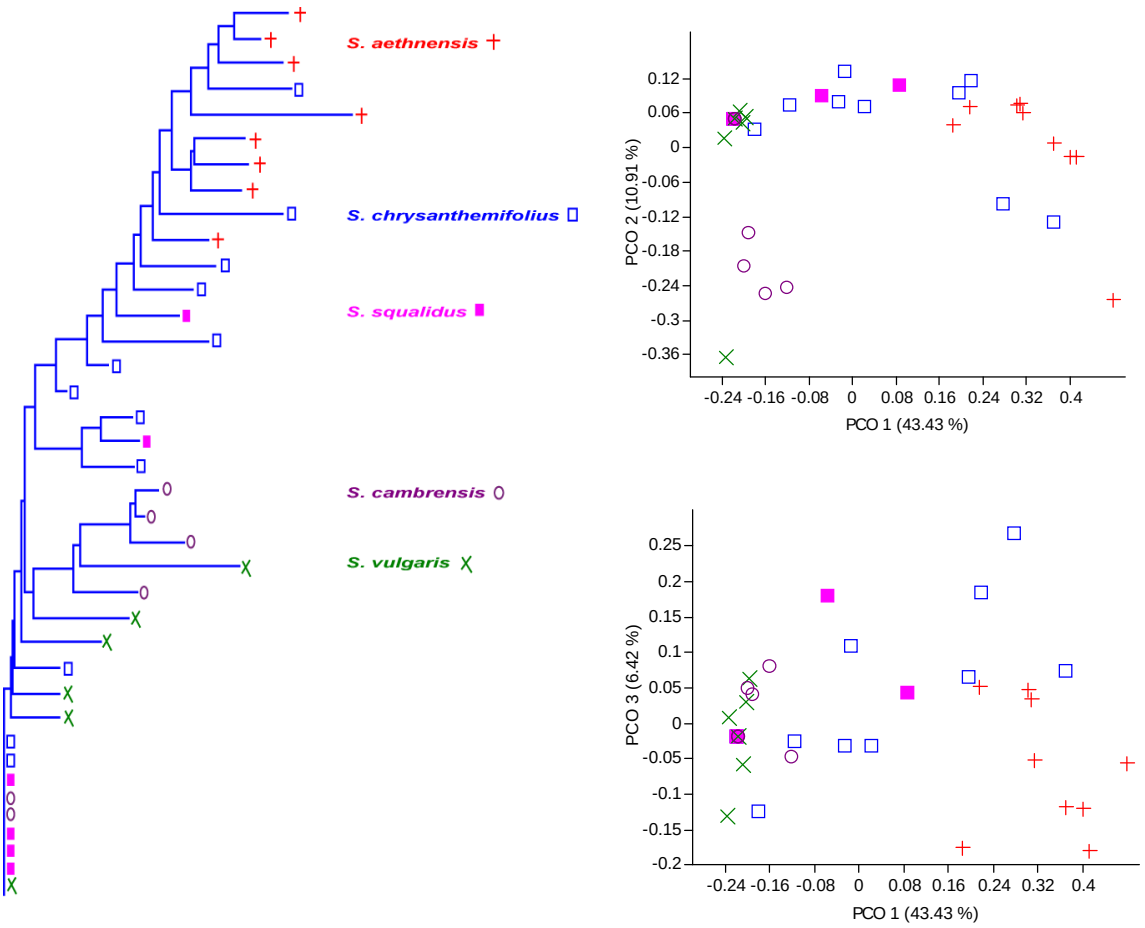


Figure A.4: NJ tree and PCO plots of *Senecio* sp. U49/snoR77Y fragment profiles. NJ and PCO analyses of 38 samples made up of 5 species are based on fragment variation (12 fds) and dice genetic similarities of the U49F/snoR77YR primer pair. The first three axis of the PCO explain 60.76 % of the variation within the dataset.

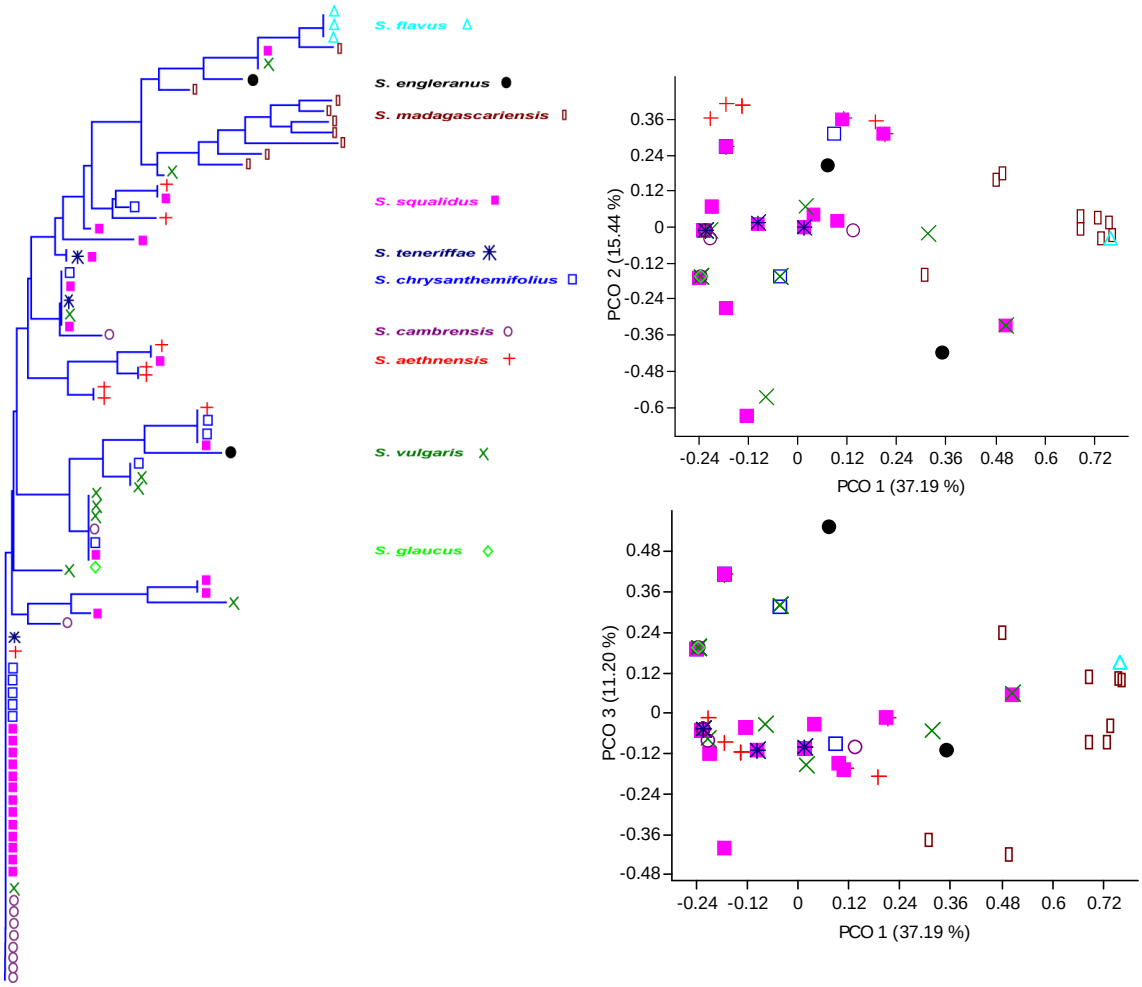


Figure A.5: NJ tree and PCO plots of *Senecio* sp. snoR13/U18 fragment profiles. NJ and PCO analyses of 88 samples made up of 10 species are based on fragment variation (15 fds) and dice genetic similarities of the snoR13F/U18R primer pair. The first three axis of the PCO explain 64.66 % of the variation within the dataset.

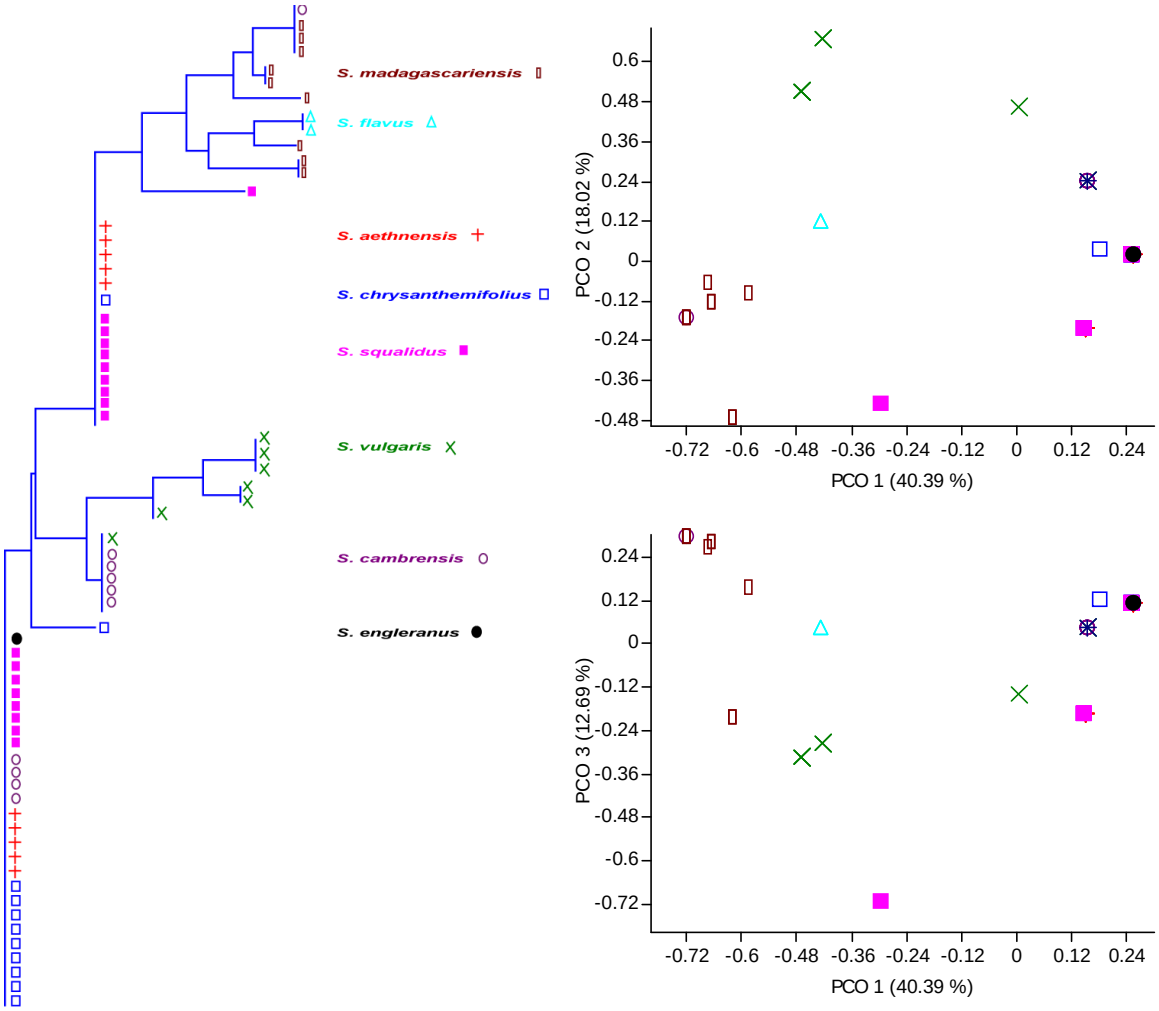


Figure A.6: NJ tree and PCO plots of *Senecio* sp. U18/U54 fragment profiles. NJ and PCO analyses of 69 samples made up of 8 species are based on fragment variation (8 fds) and dice genetic similarities of the U18F/U54R primer pair. The first three axis of the PCO explain 71.10 % of the variation within the dataset.

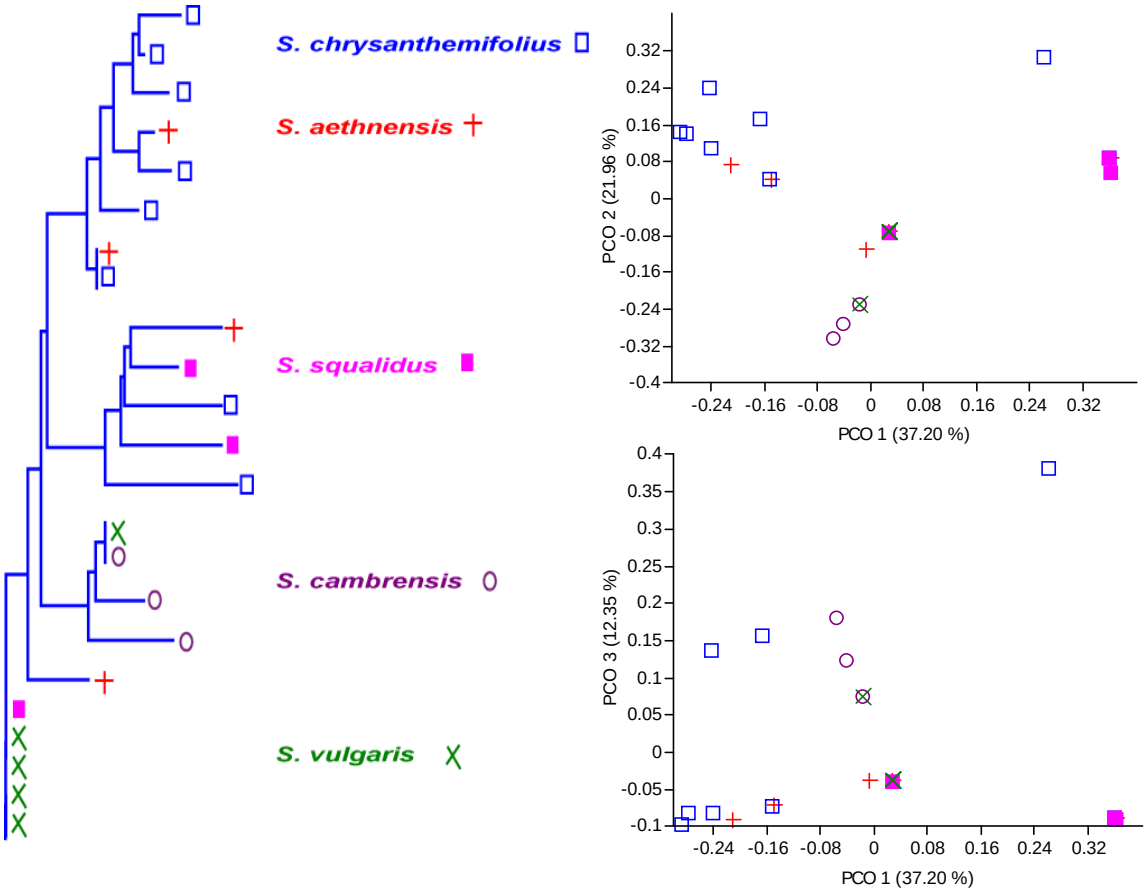


Figure A.7: NJ tree and PCO plots of *Senecio* sp. snoR13/U54 fragment profiles. NJ and PCO analyses of 24 samples made up of 5 species are based on fragment variation (16 fds) and dice genetic similarities of the snoR13F/U54R primer pair. The first three axis of the PCO explain 72.51 % of the variation within the dataset.

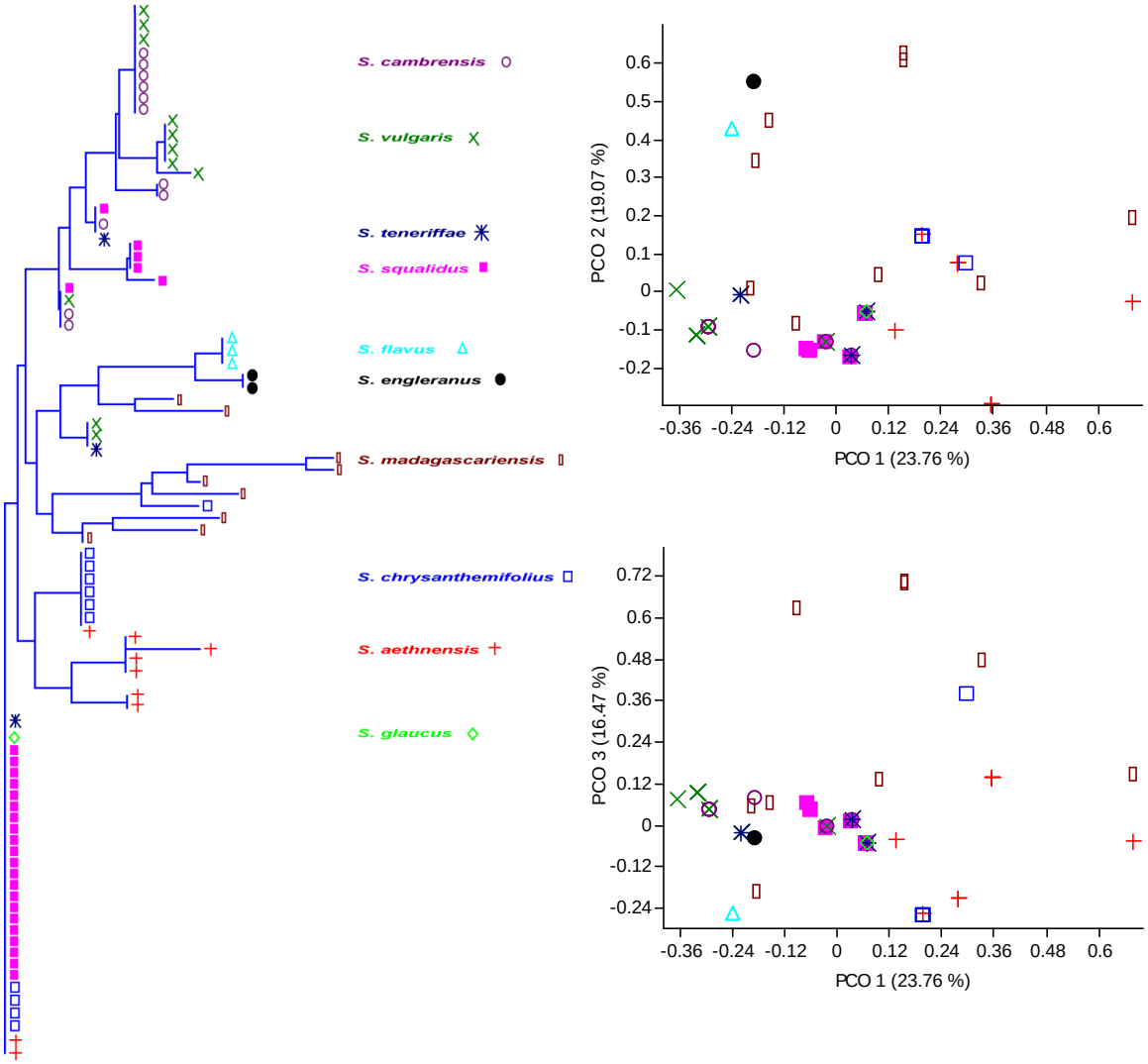


Figure A.8: NJ tree and PCO plots of *Senecio* sp. U61/snoR14 fragment profiles. NJ and PCO analyses of 88 samples made up of 10 species are based on fragment variation (14 fds) and dice genetic similarities of the Uu1F/snoR14R primer pair. The first three axis of the PCO explain 72.51 % of the variation within the dataset.

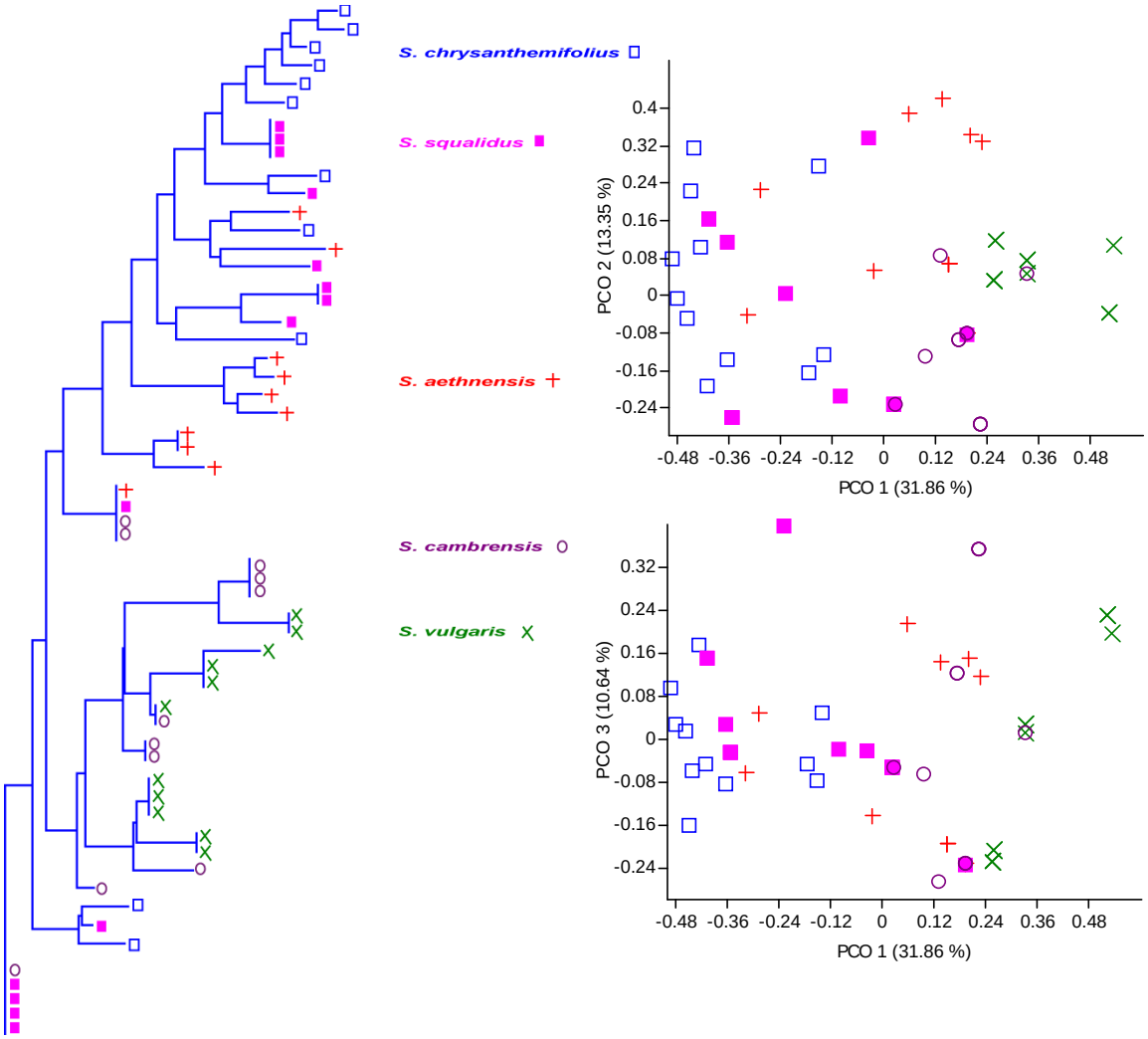


Figure A.9: NJ tree and PCO plots of *Senecio* sp. snoR37/snoR22 fragment profiles. NJ and PCO analyses of 58 samples made up of 5 species are based on fragment variation (17 fds) and dice genetic similarities of the snoR37/snoR22 primer pair. The first three axis of the PCO explain 55.85 % of the variation within the dataset.

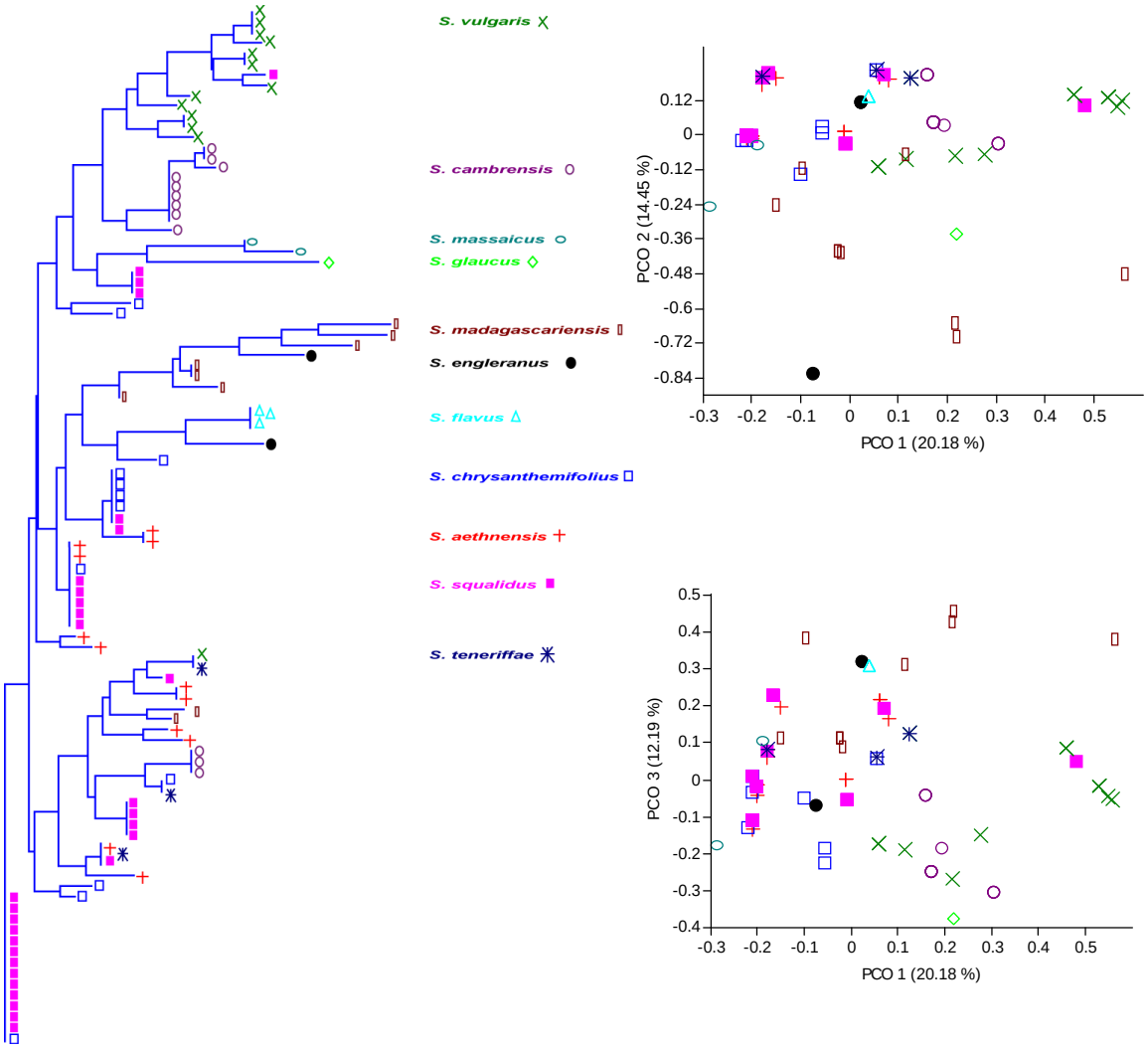


Figure A.10: NJ tree and PCO plots of *Senecio* sp. snoR22/snoR23 fragment profiles. NJ and PCO analyses of 96 samples made up of 11 species are based on fragment variation (27 fds) and dice genetic similarities of the snoR22F/snoR23R primer pair. The first three axis of the PCO explain 46.82 % of the variation within the dataset.

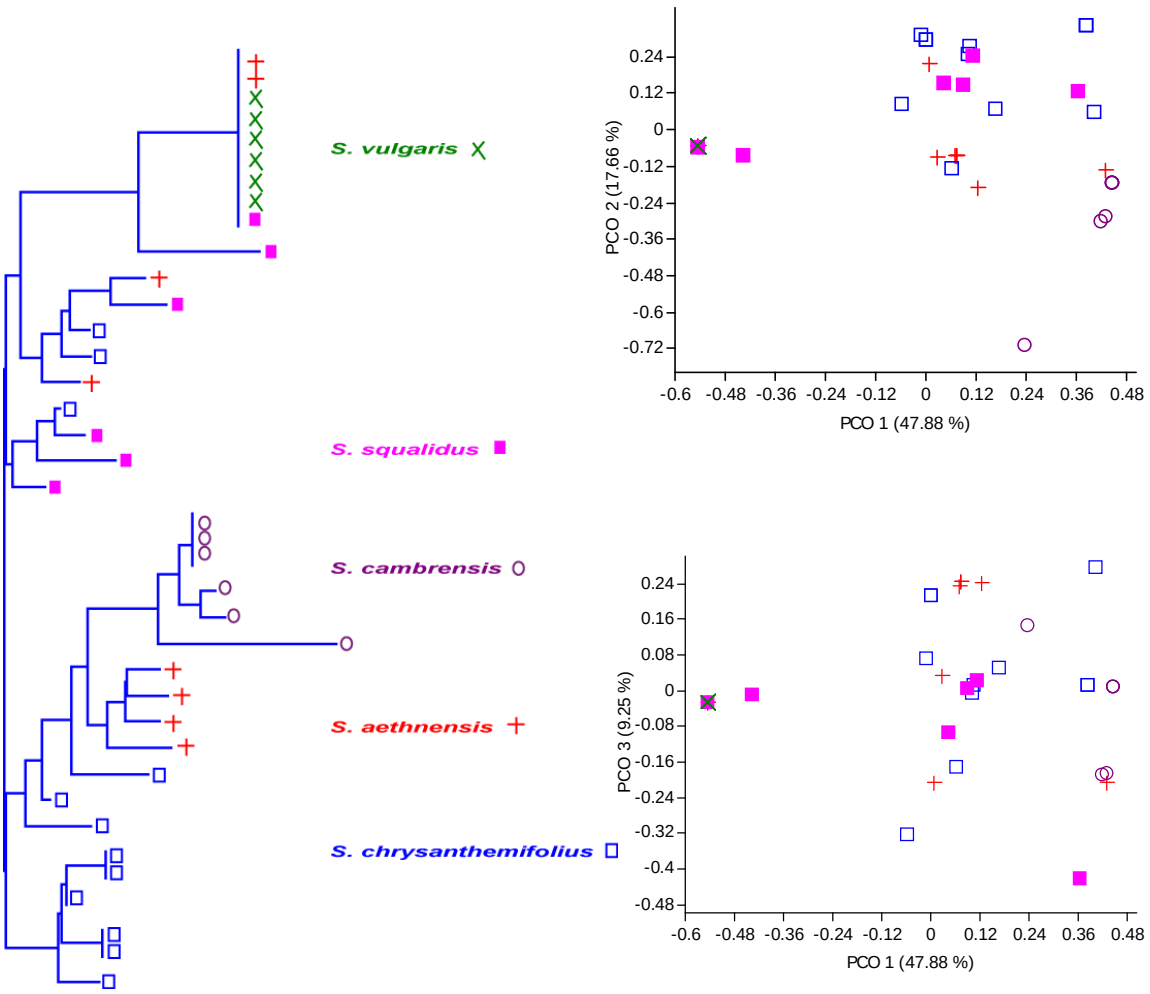


Figure A.11: NJ tree and PCO plots of *Senecio* sp. snoR37/snoR23 fragment profiles. NJ and PCO analyses of 38 samples made up of 5 species are based on fragment variation (18 fds) and dice genetic similarities of the snoR37F/snoR23R primer pair. The first three axis of the PCO explain 74.79 % of the variation within the dataset.

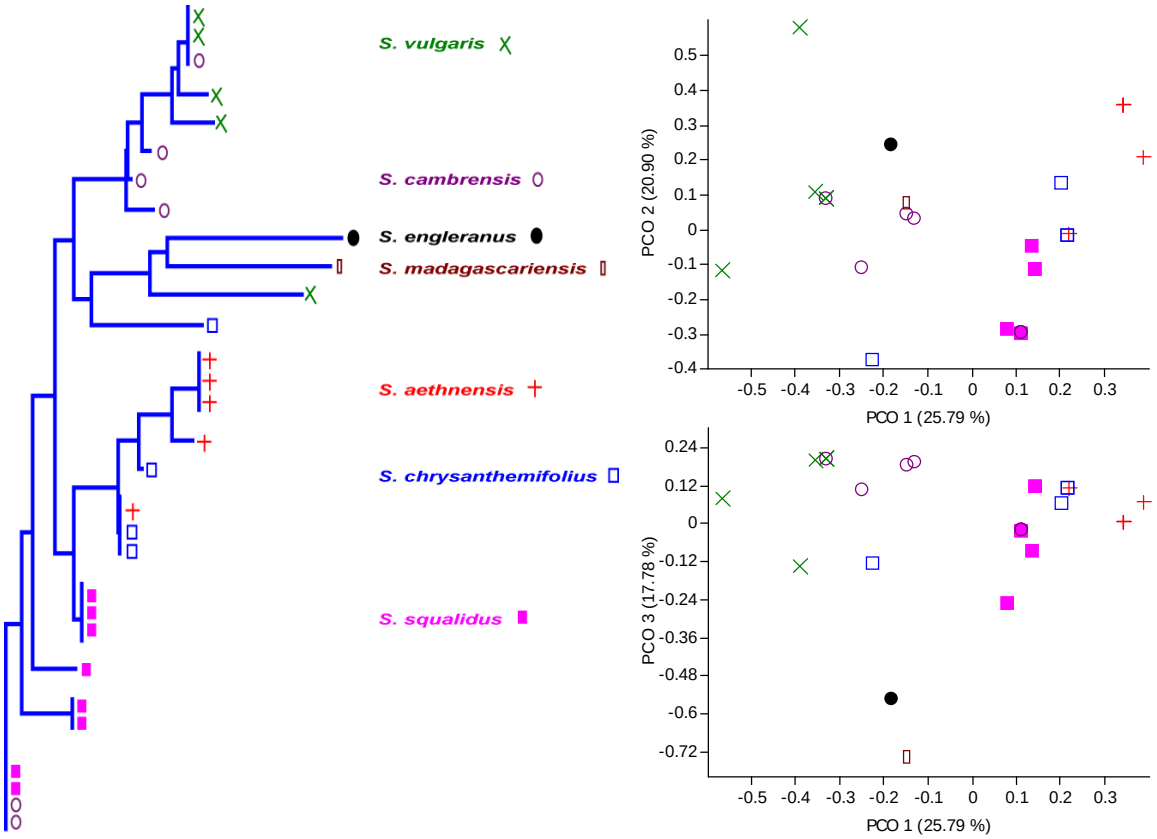


Figure A.12: NJ tree and PCO plots of *Senecio* sp. snoR66/119bR1 fragment profiles. NJ and PCO analyses of 30 samples made up of 7 species are based on fragment variation (14 fds) and dice genetic similarities of the U31F/U51R primer pair. The first three axis of the PCO explain 64.47 % of the variation within the dataset.

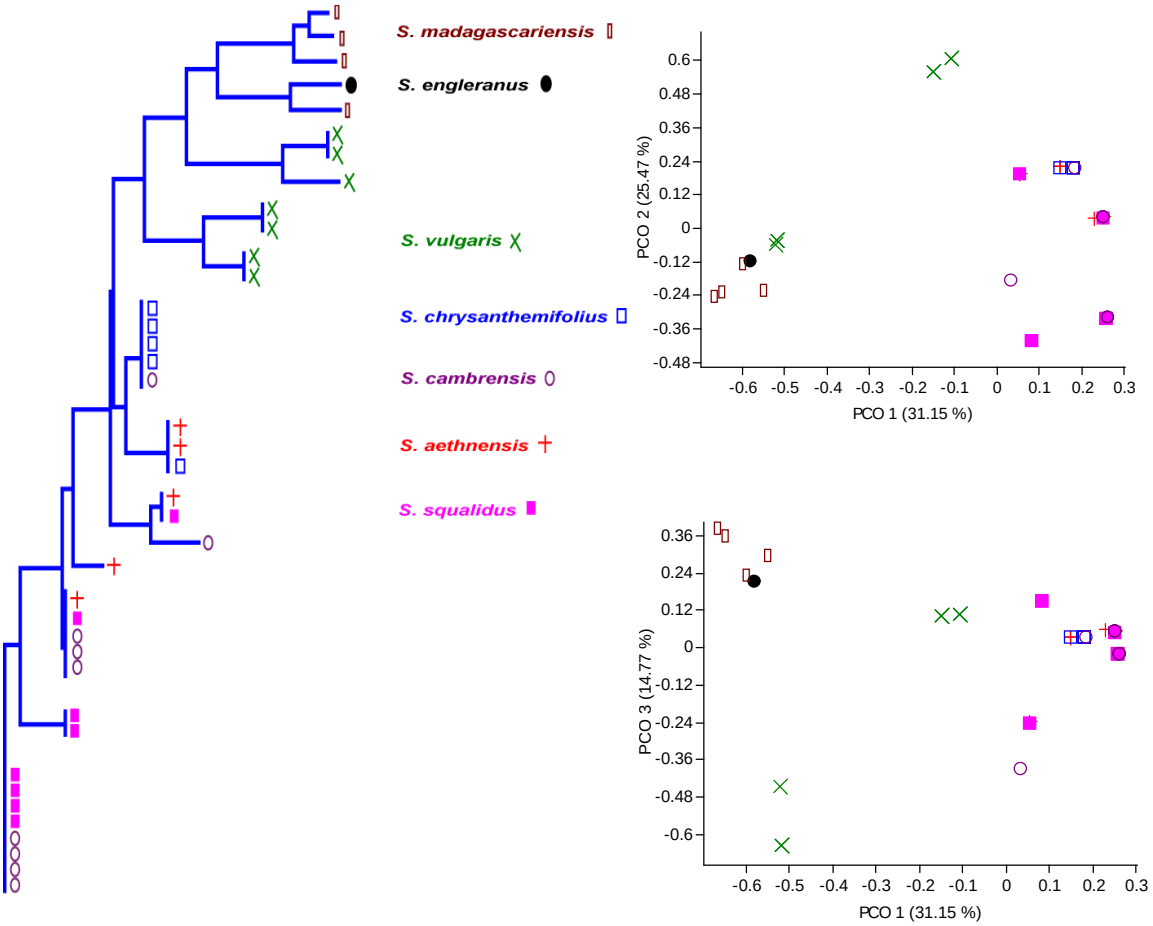


Figure A.13: NJ tree and PCO plots of *Senecio* sp. snoR66/119bR2 fragment profiles. NJ and PCO analyses of 39 samples made up of 7 species are based on fragment variation (13fds) and dice genetic similarities of the snoR66F/119bR2 primer pair. The first three axis of the PCO explain 71.39 % of the variation within the dataset.

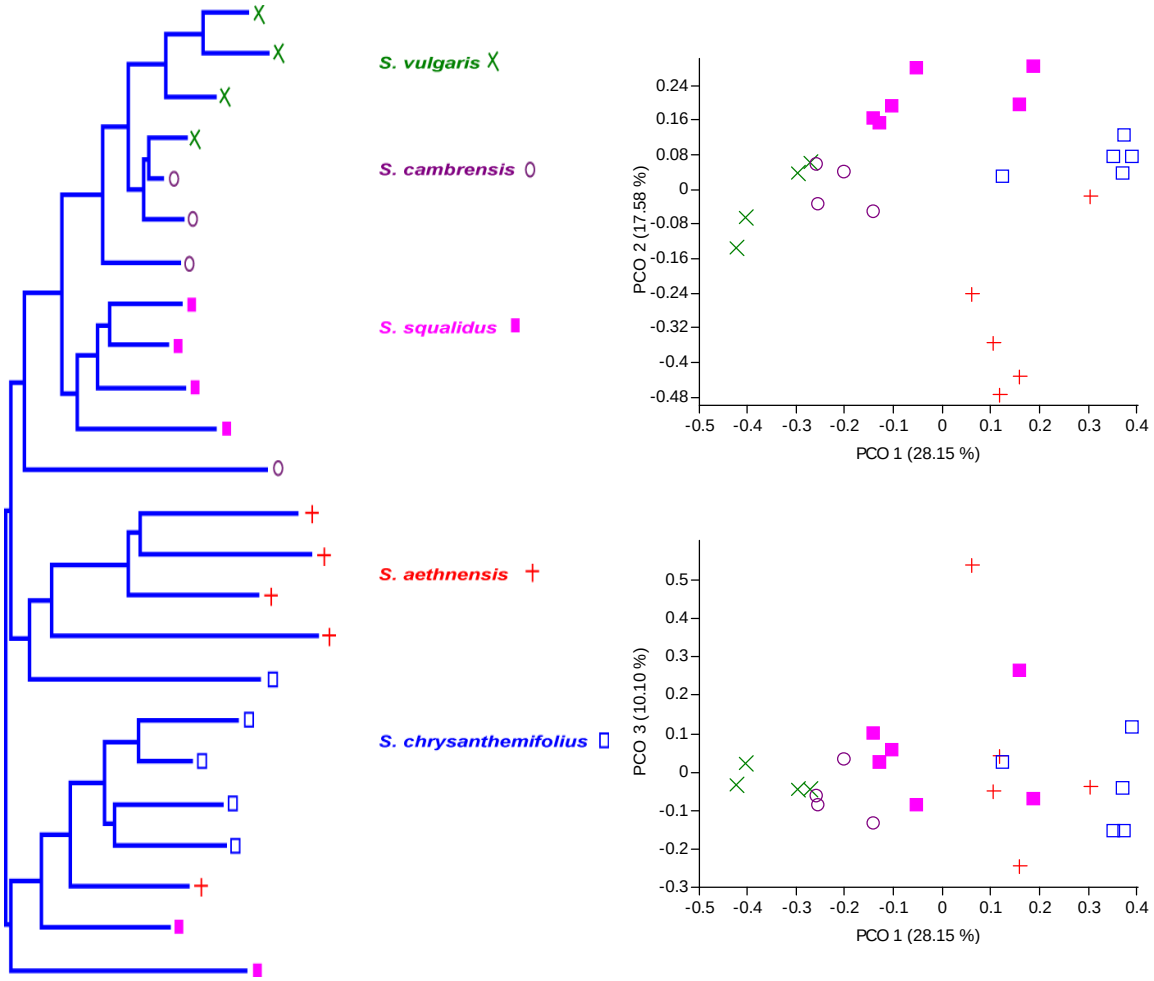


Figure A.14: NJ tree and PCO plots of *Senecio* sp. snoR114/snoR85 fragment profiles. NJ and PCO analyses of 24 samples made up of 5 species are based on fragment variation (39 fds) and dice genetic similarities of the snoR114F/snoR85R primer pair. The first three axis of the PCO explain 55.83 % of the variation within the dataset.

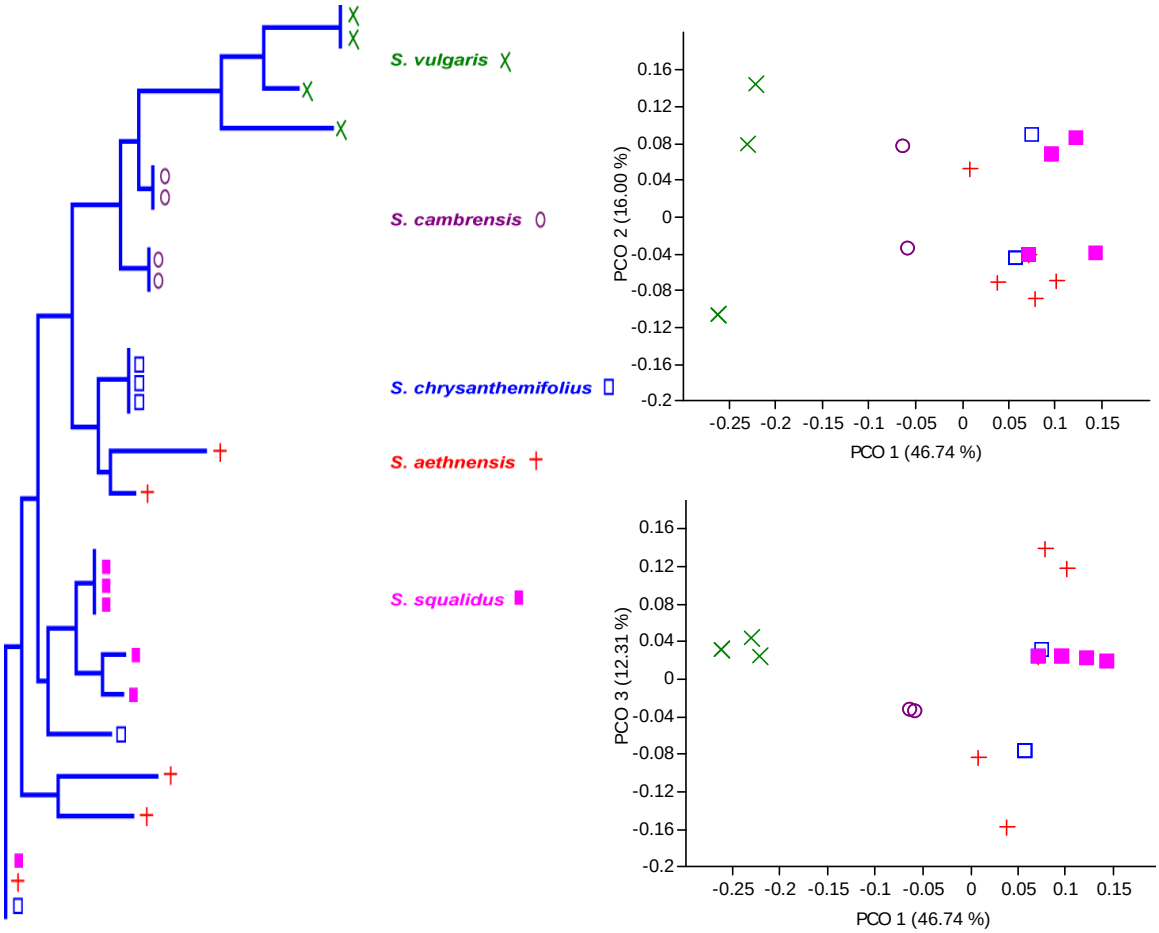


Figure A.15: NJ tree and PCO plots of *Senecio* sp. snoR115/snoR85 fragment profiles. NJ and PCO analyses of 24 samples made up of 5 species are based on fragment variation (21 fds) and dice genetic similarities of the snoR115F/snoR85R primer pair. The first three axis of the PCO explain 75.05 % of the variation within the dataset.

Chapter 5

Table A.21 to Table A.32 show the lengths fragments obtained from various species for the primer pairs of each gene cluster.

Table A.21: Fragment lengths produced by primer pairs U31/U33, U31/U51 and U33/U51 of cluster A. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental).

Species	Sequence ID	U31/U33	U31/U51	U33/U51
Arabidopsis thaliana	ePCR	324, 325	445, 455	141, 151
Brassica napus	gi|150890906	353	468	131
Brassica oleraceae	gi|150918428	263		
Raphanus raphanus ssp. *raphanus*	gi|167487179	333	481	163
Raphanus raphanus. ssp. *maritimus*	gi|166104104	333	489	171
Raphanus sativus	gi|166139508	346	495	164
Senecio sp	experimental	200, 340	192, 284, 340, 480	150

Table A.22: Fragment lengths produced by the U14-3/U14-4 primer pair of cluster B. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental).

Species	Sequence ID	U14-3/U14-4
Arabidopsis thaliana	ePCR	118, 124, 285, 321, 483, 650
Brassica napus	gi|151179762	313, 479, 122
Brassica rapa	gi|150125250	330
Citrus sinensis	gi|188372223	386, 650, 220
Glycine max	gi|22523195	202
Guizotia abyssinica	gi|211706752	291
Helianthus exilis	gi|113303713	298
Lactuca saligna	gi|83786361	169
Lactuca sativa	gi|90511348	344
Lotus japonicus	gi|93652383	190
Medicago truncatula	gi|161110397	227
Meloidogyne arenaria	gi|126164181	295
Nicotiana tabacum	gi|92012225	253
Populus trichocarpa x Populus deltoides	gi|52536034	204
Populus trichocarpa x Populus deltoides	gi|52527682	214, 462, 204
Raphanus sativus var. oleiformis	gi|156156262	273, 448, 140
Senecio vulgaris subsp. vulgaris	gi|89507118	
Vigna unguiculata	gi|182400318	204, 423, 175
Senecio sp.	experimental	129, 130, 680, 694

Table A.23: Fragment lengths produced by the U36/U38 primer pair of cluster C. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental).

Species	Sequence ID	U36/U38
Arabidopsis thaliana	ePCR	100, 115, 116
Festuca pratensis	gi|237575617	158
Helianthus annuus	gi|211618741	92
Helianthus ciliaris	gi|125409402	92
Ipomoea nil	gi|74394972	135
Limnanthes alba	gi|166358285	113
Mesembryanthemum crystallinum	gi|8578441	98
Raphanus raphanistrum subsp. *maritimus*	gi|161516828	123
Triphysaria pusilla	gi|159779093	131
Zea mays	gi|211347758	141
Senecio sp	experimental	94, 157

Table A.24: Fragment lengths produced by primer pairs U49/snoR2, U49/snoR77Y and snoR2/snoR77Y of cluster D. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental).

Species	Sequence ID	U49/snoR2	U49/snoR77Y	snoR2/snoR77Y
Arabidopsis thaliana	ePCR	249, 331, 352	452, 462	145, 156
Barnadesia spinosa	gi|211684238	336		
Cichorium intybus	gi|124611070	324	480	175
Helianthus ciliaris	gi|125404409	339	453	132
Helianthus paradoxus	gi|125472412	312	429	135
Lactuca virosa	gi|84036229	316		
Medicago truncatula	gi|20455926	297	446	174
Senecio vulgaris	gi|89504909	596		
Solanum habrochaites	gi|261475319	299	422	146
Solanum melongena	gi|261672523	305	444	164
Zinnia violacea	gi|41119507			181
Senecio sp.	experimental	120, 123, 127, 131, 134, 155, 216, 235	120, 122, 127, 131, 134, 303 308, 391, 553	157

Table A.25: Fragment lengths produced by primer pairs snoR13/U18, snoR13/U54 and U18/U54 of cluster D. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	SR13/U18	SR13/U54	U18/U54
Arabidopsis thaliana	ePCR	126 (2x)	283	174
Elymus lanceolatus	gi|207458184	103		
Euphorbia esula	gi|76858228	102		
Glycine max	gi|254313602	178	332	171
Hordeum vulgare	gi|24294814	115		
Malus x domestica	gi|91029133	157		
Medicago truncatula	gi|30099477	136	290	171
Mimulus guttatus	gi|238369891	175		
Nicotiana tabacum	gi|254649847	138	306	185
Raphanus raphanistrum subsp. *raphanistrum*	gi|154182120	139		
Solanum lycopersicum	gi|225416350	144	275	148
Trifolium pratense	gi|86098512	151	311	177
Triticum aestivum	gi|25228716	105		
Senecio sp	experimental	120	368, 379, 645	160, 165

Table A.26: Fragment lengths produced by the U61/snoR14 primer pair of cluster F. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	U61/SR14
Allium cepa	gi|34470953	240
Arabidopsis thaliana	ePCR	147
Cicer arietinum	gi|241795628	173
Cichorium endivia	gi|125355664	328
Citrullus lanatus	gi|198407024	171
Gossypium hirsutum	gi|73848720	159
Lotus japonicus	gi|179640647	182
Manihot esculenta	gi|164384573	137
Manihot esculenta	gi|164388849	150
Populus nigra	gi|161926056	122
Populus trichocarpa	gi|73885771	114
Solanum lycopersicum	gi|9456251	99
Solanum tuberosum	gi|78747558	159
Vitis vinifera	gi|110425778	188
Senecio sp.	experimental	116, 120, 123, 132

Table A.27: Fragment lengths produced by the snoR29/snoR30 primer pair of cluster G. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	SR29/SR30
Aquilegia formosa x Aquilegia pubescens	gi|75452514	217
Arabidopsis thaliana	ePCR	284
Brassica napus	gi|151038092	272
Carthamus tinctorius	gi|125387689	306
Citrus sinensis	gi|188455503	225
Glycine max	gi|213588053	266
Gossypium hirsutum	gi|84173523	224
Lactuca virosa	gi|84014816	227
Manihot esculenta	gi|164397641	280
Medicago truncatula	gi|11934437	209
Panicum virgatum	gi|197953775	249
Papaver somniferum	gi|189456616	227
Populus trichocarpa	gi|73894156	394
Solanum lycopersicum	gi|4382859	187
Solanum tuberosum	gi|52619648	182, 213, 536
Sorghum bicolor	gi|45988116	256
Taraxacum kok-saghyz	gi|68257999	235
Theobroma cacao	gi|215536047	231
Zea mays	gi|213175207	260
Senecio sp.	experimental	197, 203, 209, 213, 216, 231

Table A.28: Fragment lengths produced by the U80-1/U80-2 primer pair of cluster H. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental).

Species	Sequence ID	U80-1/U80-2
Actinidia chinensis	gi|195250959	63
Arabidopsis thaliana	ePCR	61, 66
Barnadesia spinosa	gi|211684399	63
Brassica napus	gi|126473296	58
Brassica rapa subsp. pekinensis	gi|179828540	61
Carica papaya	gi|186875731	58
Carthamus tinctorius	gi|125365806	67
Citrus sinensis	gi|188446220	60
Festuca pratensis	gi|237606146	61
Festuca pratensis	gi|237607495	62, 61, 517
Glycine max	gi|7591490	60
Gossypium hirsutum	gi|164335562	63
Guizotia abyssinica	gi|211707946	59
Juglans hindsii x Juglans regia	gi|133868740	58
Lactuca sativa	gi|90512182	59
Nicotiana tabacum	gi|224705621	62
Oryza sativa	gi|29646057	70
Panicum virgatum	gi|254543715	69
Populus trichocarpa x Populus deltoides	gi|73936044	63
Raphanus raphanistrum subsp. *raphanistrum*	gi|154180020	57
Salmo salar	gi|117436796	59
Saprolegnia parasitica	gi|76445167	62
Solanum lycopersicum	gi|9429768	66
Vitis vinifera	gi|110365818	61
Zea mays	gi|32913295	70
Senecio sp.	experimental	56, 138, 286

Table A.29: Fragment lengths produced by the U15/snoR7 primer pair of cluster I. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	U15/SR7
Arabidopsis thaliana	ePCR	143, 183, 294
Glycine max	gi|192297583	188
Oryza sativa	gi|117227779	161, 384
Populus trichocarpa	gi|24073220	197
Raphanus raphanistrum subsp. *raphanistrum*	gi|166120078	164, 369
Saccharum hybrid	gi|268805515	142
Triphysaria pusilla	gi|159683902	216
Triticum aestivum	gi|9445038	153, 314
Tropaeolum majus	gi|215785250	159
Vitis vinifera	gi|27579784	132
Zea mays	gi|76019688	175
Senecio sp.	experimental	157

Table A.30: Fragment lengths produced by primer pairs snoR37/snoR22, snoR37/snoR23 and snoR22/snoR23 of cluster J. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR

Species	Sequence ID	SR37/SR22	SR37/SR23	SR22/SR23
Arabidopsis thaliana	ePCR	166, 212, 213, 389	362, 412, 588	217, 220 (2x), 397
Barnadesia spinosa	gi|211666183	220	442	243
Brassica napus	gi|150155151	178		
Citrus sinensis	gi|188232298	288	510	243
Euphorbia esula	gi|76860179	150	339	210
Gossypium hirsutum	gi|109869959	197	434	156
Lactuca sativa	gi|22234984	195	385	211
Lactuca serriola	gi|83917712	214	426	233
Medicago truncatula	gi|260527072			256
Phaseolus vulgaris	gi|171544740			320
Populus tremula var. glandulosa	gi|57890243	200		
Raphanus raphanistrum subsp. *maritimus*	gi|166100767	191	399	229
Raphanus raphanistrum subsp. *maritimus*	gi|167475249			234
Raphanus sativus	gi|167451966	188		
Raphanus sativus var. *oleiformis*	gi|166134883			236
Solanum lycopersicum	gi|225415904	215	434	240
Senecio sp.	experimental	105, 120, 126, 136, 142, 150, 330, 343,	310, 333, 355, 530, 548, 560	202, 219, 224, 440

Table A.31: Fragment lengths produced by primer pairs snoR66/119b1, snoR66/119b2 of cluster M. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	SR66/119b1	SR66/119b2
Arabidopsis thaliana	ePCR	162	219
Cyamopsis tetragonoloba	gi|117894741	168	226
Glycine soja	gi|26044093	192	240
Helianthus tuberosus	gi|125445895	155	212
Lactuca perennis	gi|83886679	180	234
Lactuca sativa	gi|83992923	167	221
Lactuca serriola	gi|22438909	167	221
Lactuca virosa	gi|84009164	178	232
Medicago truncatula	gi|13780193	165	206
Populus tremula x Populus tremuloides	gi|24076600	186	242
Populus trichocarpa	gi|52386830	216	
Raphanus raphanistrum subsp. *landra*	gi|166125710	157	214
Solanum tuberosum	gi|10448481		127
Solanum tuberosum	gi|20170484		127
Solanum tuberosum	gi|21915632		127
Vitis vinifera	gi|110390903	174	207
Senecio sp.	experimental	100, 105, 108 125, 210	160, 165 168, 257

Table A.32: Fragment lengths produced by primer pairs snoR114/snoR85 and snoR115/snoR85 of cluster N. The length of fragments was obtained by either ePCR, calculation using ESTs (gi numbers) or fragment analysis (experimental). SR = snoR.

Species	Sequence ID	SR114/SR85	SR115/SR85
Arabidopsis thaliana	ePCR	255, 438	146, 329
Euphorbia esula	gi|76858228	405	290
Malus x domestica	gi|48110246		203, 367
Populus trichocarpa	gi|24069242	315	152
Senecio sp.	experimental	97, 123, 128 257, 438, 725, 770	97, 223, 232, 258, 354, 359, 411, 587, 610, 621, 739, 770

Figure A.16 to Figure A.24 show the gene organization of all clusters but cluster H, M and N. Cluster H was examined by one gene (U80) and, thus, shows only variation in the gene region. The gene organization of cluster M and N is shown in Chapter 3 (clusters D and E therein).

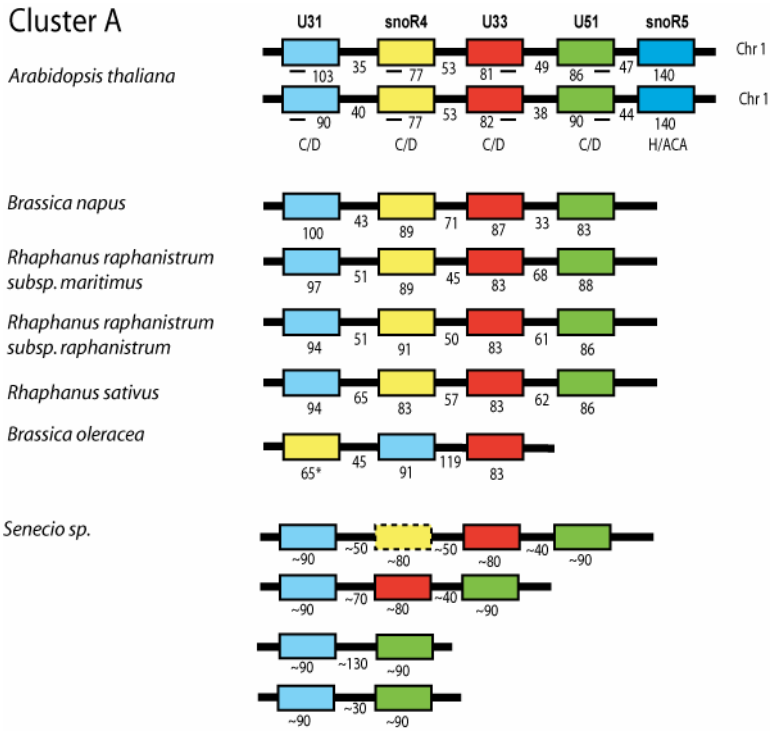


Figure A.16: Gene organization of cluster A in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. Dotted box represent a putative gene which lack supporting fragment pattern. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

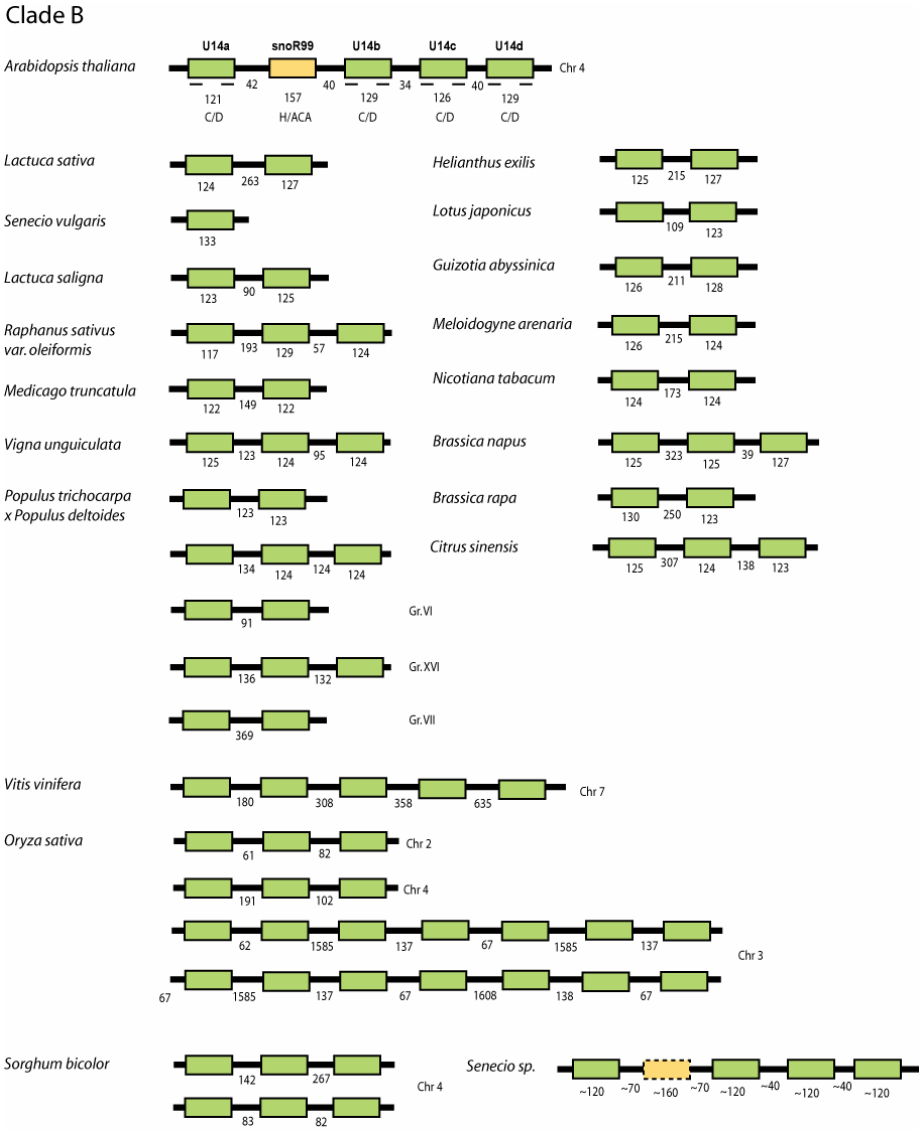


Figure A.17: Gene organization of cluster B in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequence) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) and group (Gr.) numbers for genomic sequences are shown. Note that the gene cluster of *Oryza sativa* on chromosome 3 consists of 14 U14 genes and is displayed in two lines. The approximate location of the universal primer sites in *A. thaliana* are indicated by black lines below the genes. Also note that the sequences were

only examined for U14 genes and long intergenic regions might, thus, harbour additional genes

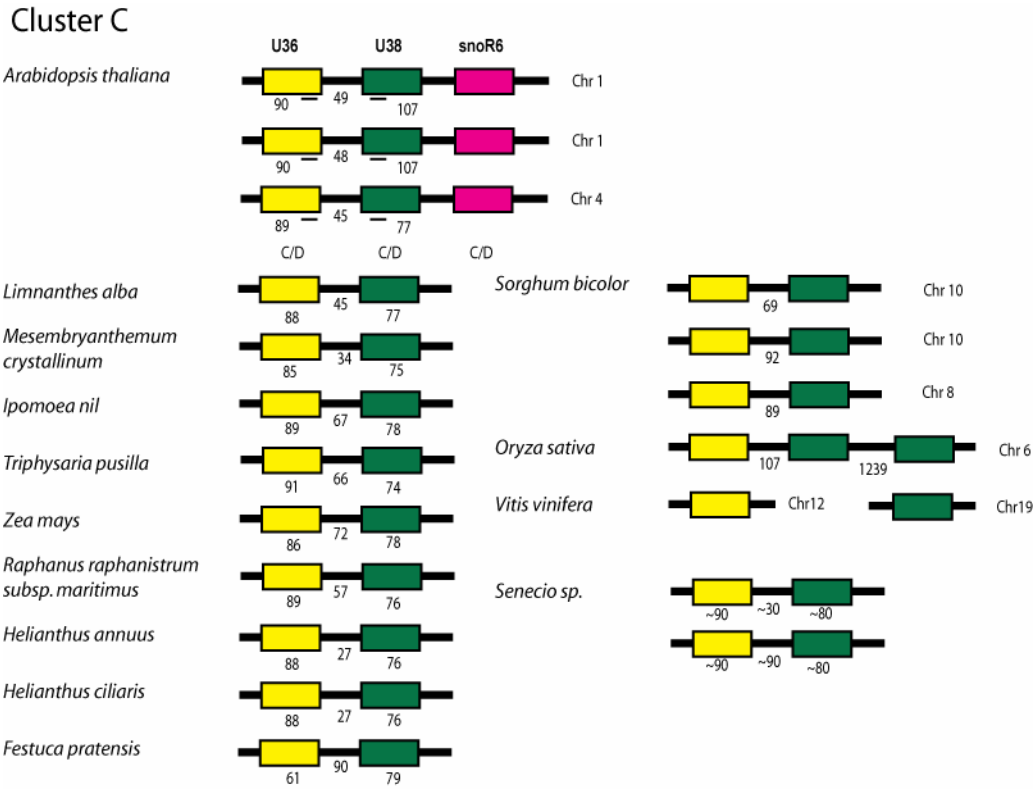


Figure A.18: Gene organization of cluster C in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

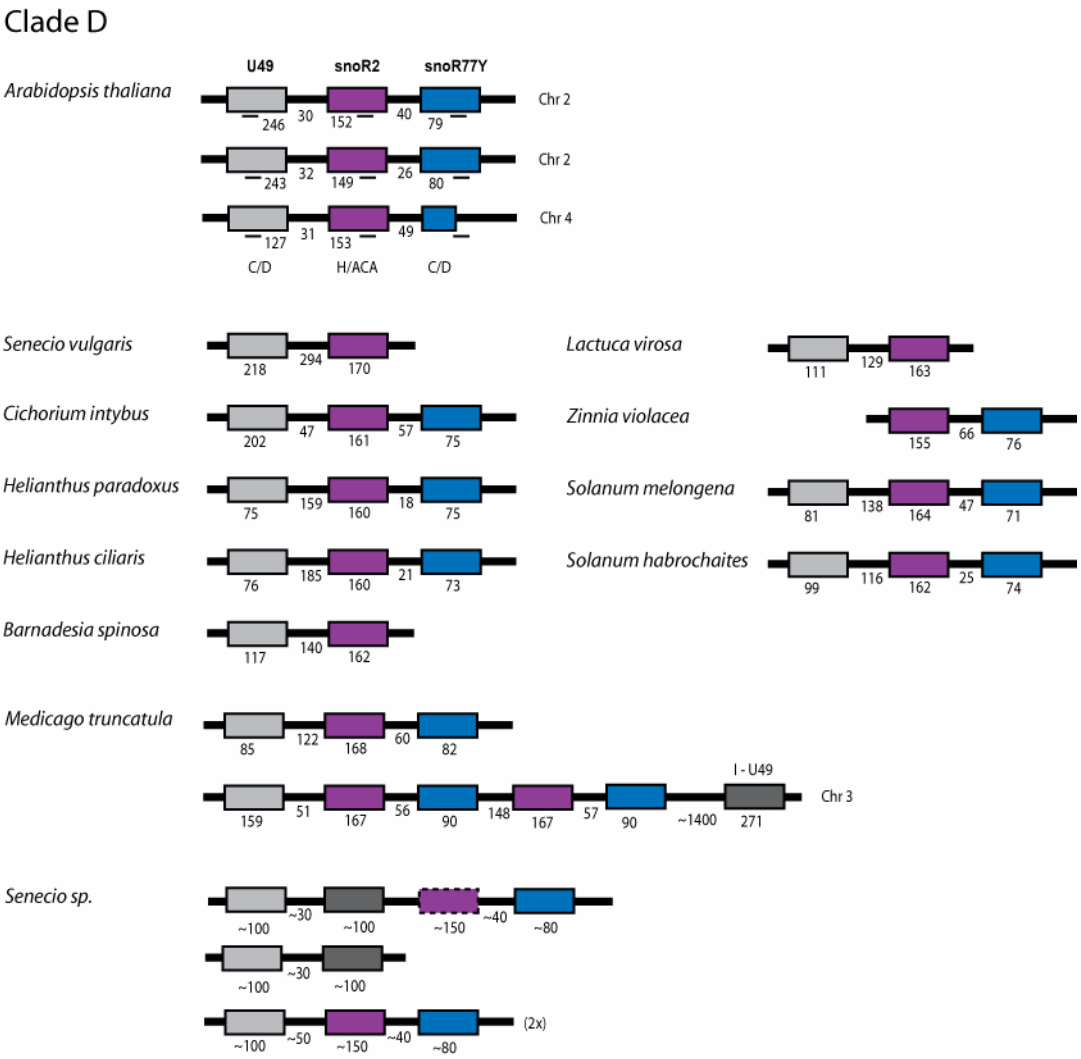


Figure A.19: Gene organization of cluster D in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. Dotted box represents a putative gene which lacks supporting fragment pattern. I-U49 = inverted U49 gene. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

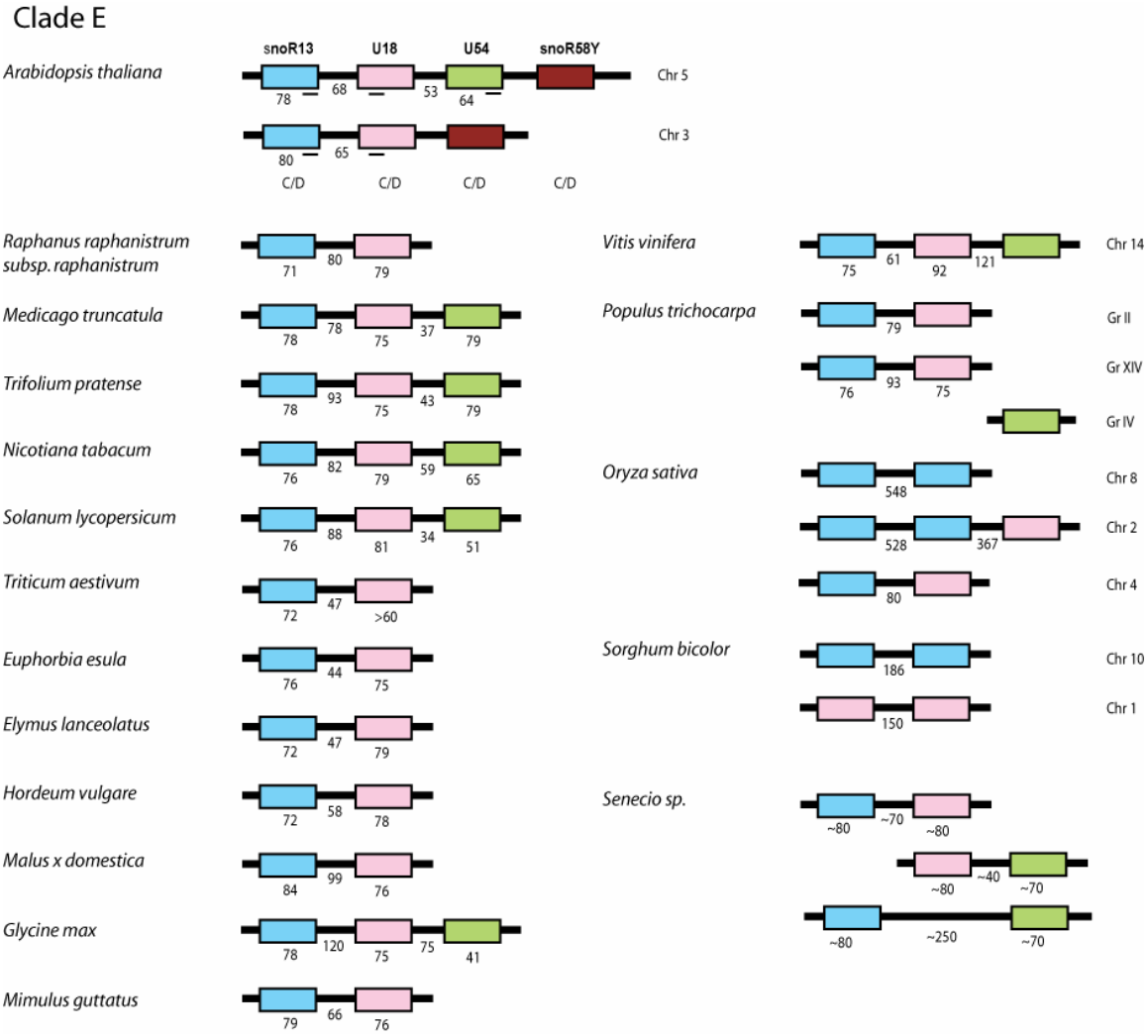


Figure A.20: Gene organization of cluster E in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

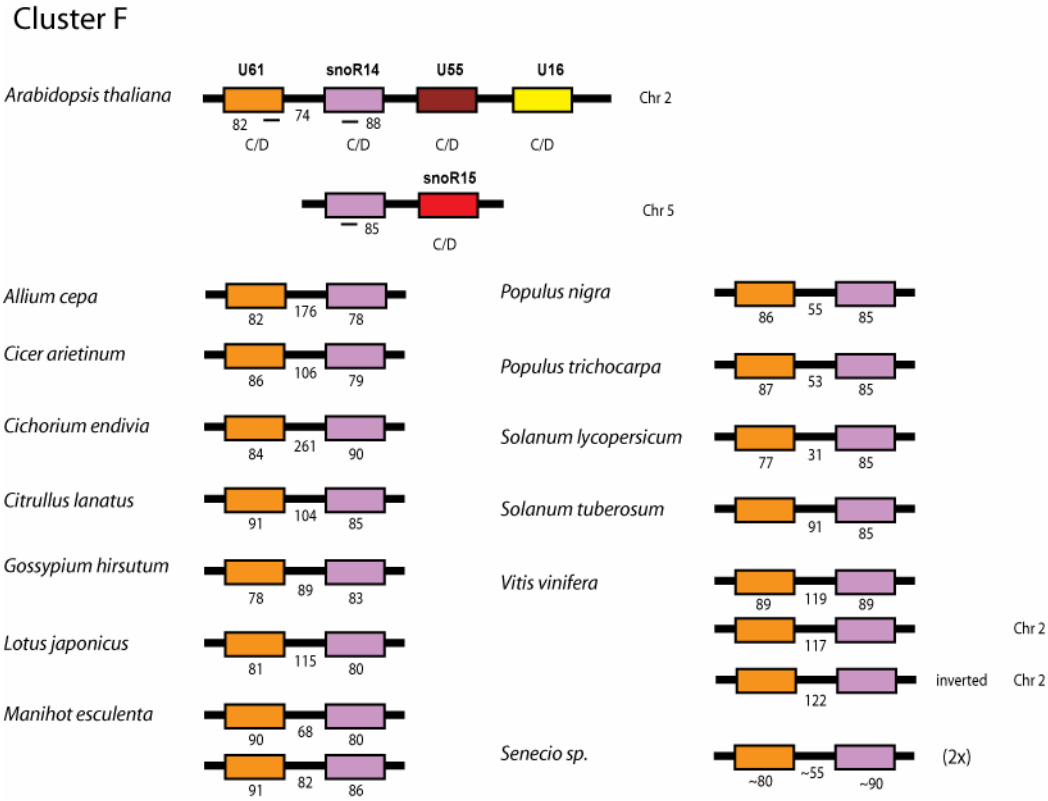


Figure A.21: Gene organization of cluster F in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

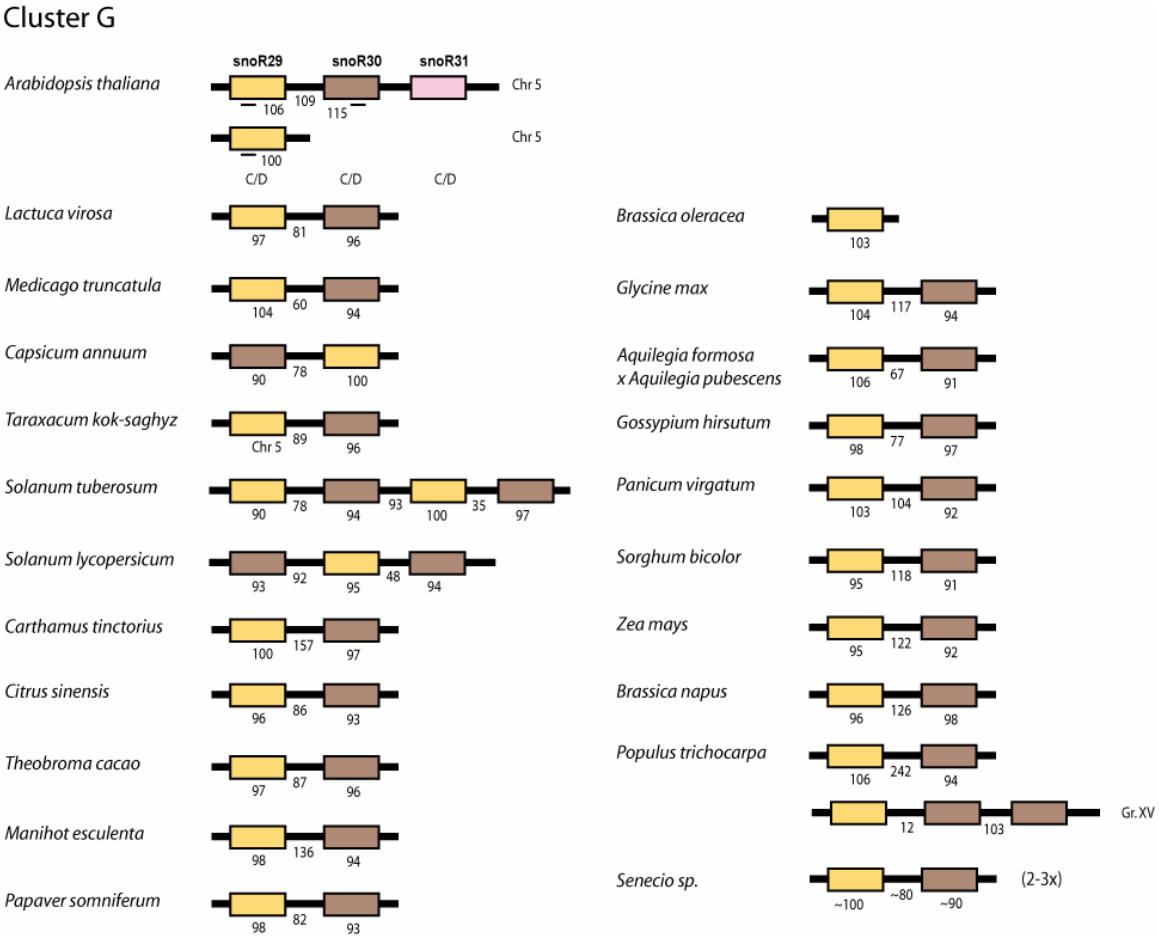


Figure A.22: Gene organization of cluster G in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

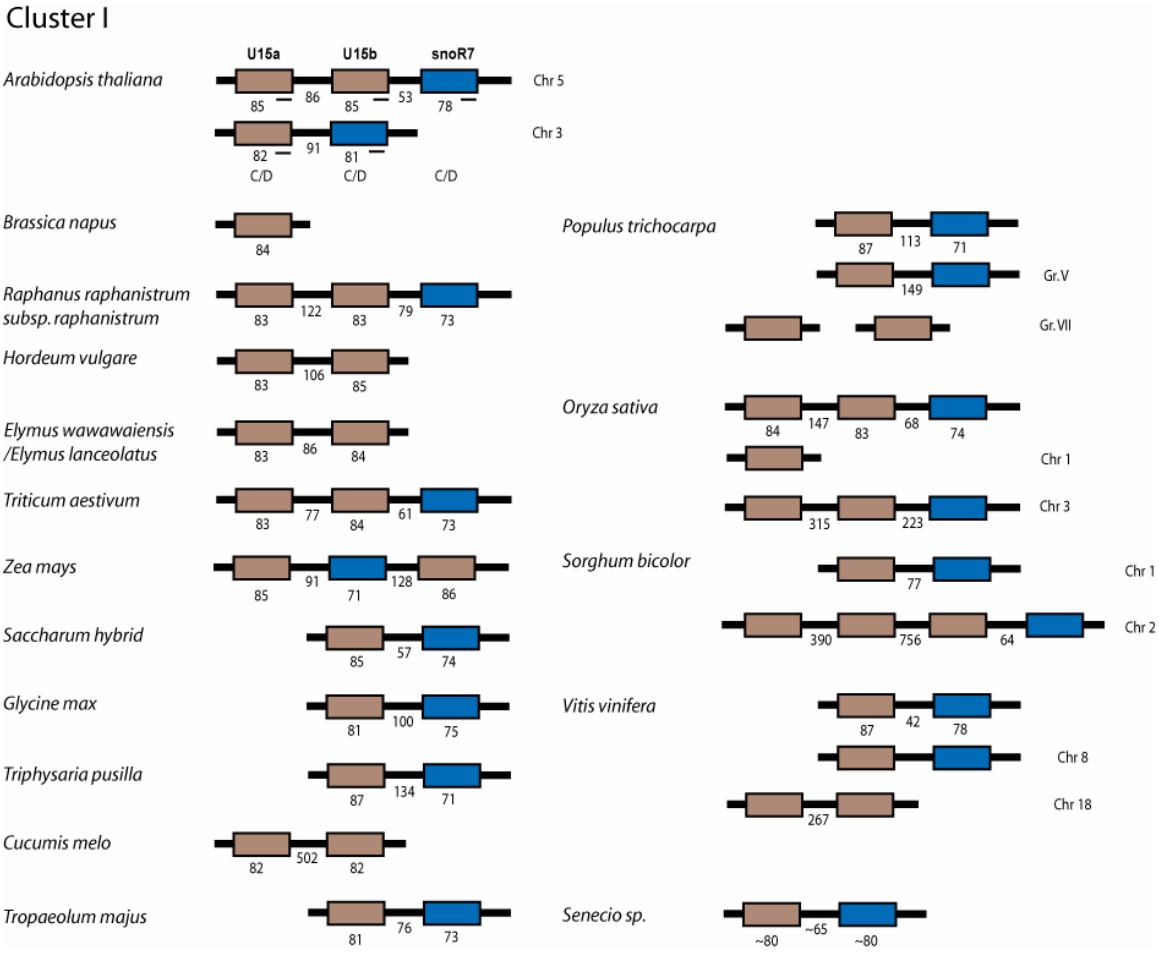


Figure A.23: Gene organization of cluster I in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.

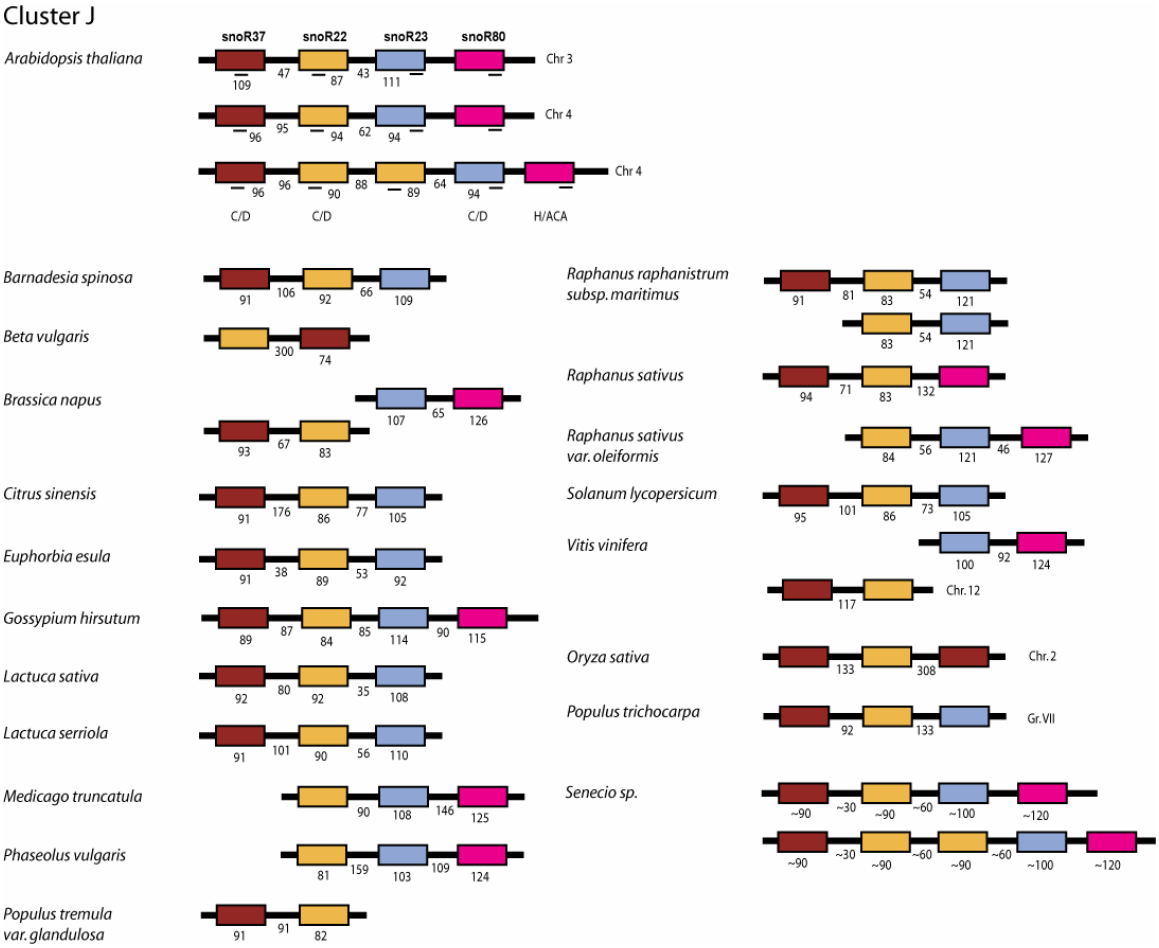


Figure A.24: Gene organization of cluster J in various species. Boxes represent gene sequences with different genes within a cluster indicated by different colours. The names of the genes are given above the boxes and the lengths of genes (only for ESTs and *Arabidopsis* sequences) and intergenic regions are indicated by numbers below the boxes and the lines, respectively. The reconstructed gene cluster for *Senecio* is also shown. The chromosome (chr.) numbers for genomic sequences are shown. The approximate location of the universal primer sites in Arabidopsis thaliana are indicated by black lines below the genes.