Adaptive metabolic gene clusters as toolkits for chemical innovation: Investigation of the origin of the avenacin gene cluster for synthesis of defense compounds in oats

Hoi Yee Chu (Athena)

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

University of East Anglia
John Innes Centre
Norwich, the United Kingdoms
September 2013

Abstract

Operon-like gene clusters are functional cassettes of physically linked and non-homologous genes involved in the same pathway. To date, 20 such plant gene clusters have been discovered, all of which are involved in specialised metabolism. Plant gene clusters raise interesting biological questions about their importance and the drivers behind their formation. This thesis describes the investigation of the evolution of the avenacin gene cluster, discovered in the diploid oat *Avena strigosa* S75, via wet-bench experiments and bioinformatic analyses, commencing with the general introduction (Chapter 1).

Chapter 2 to 4 describe the survey on the avenacin production, expression pattern and phylogeny of the five characterized avenacin biosynthetic (*Sad*) genes within *Aveninae*, focusing on *Avena L.* The genomes of all *Avena spp.* investigated, including the avenacin deficient *A. longiglumis*, possess the five *Sad* gene homologues. The expression pattern of the *Sad* gene homolgoues vary in a genome-type dependent manner that it is root-specific amongst A genome oats. However, the C genome oats show root and leaf expressions, contributed by differentially expressed *Sad* gene duplicates.

Chapter 5 and 6 describe the molecular evolutionary analysis of the five gene families implicated in triterpene biosynthesis: oxidosqualene cyclases, cytochromes P450 51s, Clade 1A serine carboxypeptidase-like acyltransferases, Class I *O*-methyl transferases and Group L glycosyltransferases in monocots. Phylogenetics analyses show that these gene families evolve via duplication-neofunctionalisation, facilitated by gene GC content and exon-intron structures changes under purifying selection on amino acid sequences. Syntenic study of the triterpene biosynthetic gene families reveals the ancestral triterpene biosynthetic *OSC/CYP51* gene pair found in the ρ-WGD event.

Finally, the evolutionary model of the avenacin biosynthesis and the potential applications of the knowledge of gene clustering in systematic and synthetic biology are descibed in Chapter 7.

Acknowledgement

I am deeply in debt to my supervisor, Professor Anne Osbourn and Dr Jo Dicks, for carrying me through my PhD study with their help, supports and mentoring. It has been a great PhD experience and I would like to express my profound gratitude to both of my supervisors. I would also like to thank Professor Nick Brewin and Professor Mike Merrick for taking me on broad to the rotational PhD project, rendering me an invaluable opportunity to carry out research in plants and build my own PhD project with my supervisors.

I would like to thank all the lab members in the Osbourn lab for all their advice, accompany and support, especially Dr Rachel Melton, who gives me plentiful of technical supports, experimental advices and great patience. I would like to thank as well my supervisory team member, my rotation project mentor and rotational project supervisors, Dr Cristobal Uauy, Professor Sophien Kamoun and Dr Ane Sesma for their good advice and support. I would also like to thank Dr Paul O′Maille for the inspirations in fitness landscapes and catalytic spaces, Dr Lionel Hill for sharing his expertise in metabolite analysis and maths, Dr Tim Langdon and Professor Pilar Catalan for their kind offering of oat seeds and plants for my experiments.

I would like to thank the PhD student community here that have been making the student life pleasant and caring.

Finally, I would like to thank my family that have been very encouraging and supportive, giving me the freedom to pursue my interests and dreams, over all these years overseas.

Contents

List of Figures

Chapter 5 - Phylogenetic studies of *Sad* genes

Chapter 6 - Investigating the evolution of the avenacin gene cluster

List of Tables

Chapter 1 - General introduction

1.1. Specialised metabolism in plants

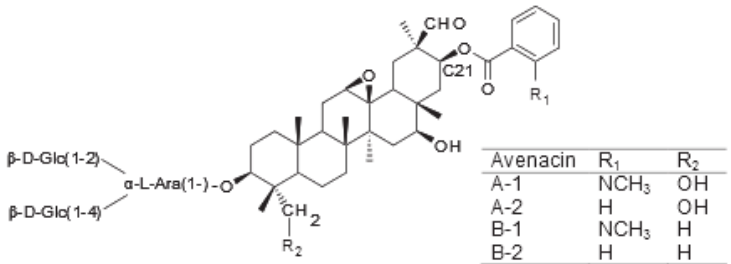
Plants produce a vast amount of specialized metabolites with tremendous variety in both structural and chemical properties (Milo and Last, 2012; Weng et al., 2012). One of these specialized metabolite classes is the terpenoids, which are involved in plant-environment interactions (Gershenzon and Dudareva, 2007; Pichersky and Lewinsohn, 2011). Triterpenes are one of the most diverse classes of chemicals produced by plants (Kliebenstein and Osbourn, 2012). Structural diversity in triterpene metabolism is achieved through the folding of the substrate 2,3-oxidosqualene into numerous and diverse conformations by enzymes known as oxidosqualene cyclases (OSC), followed by modifications of these triterpene scaffolds by combinations of tailoring enzymes such as cytochromes P450 (CYP), acyltransferases (AT) and sugar transferases (UGT) (Phillips et al., 2006; Sawai and Saito, 2011; Weng et al., 2012). For example, *Arabidopsis thaliana* has 13 OSCs, 246 CYPs and 112 UGTs, potentially generating thousands of different terpenoids in a mix-and-match manner (Sawai and Saito, 2011).

1.2. Avenacins – structure and biosynthesis

Avenacins are antifungal triterpene glycosides (saponins) that are produced by oat (*Avena species*) (Crombie and Crombie, 1986). Avenacins have been known as effective antifungal saponins since 1964 (Maizel et al., 1964). The ability to synthesise these compounds has been reported to be unique to the genus *Avena* (Crombie and Crombie, 1986). The natural variant *A. longiglumis*, which is avenacin deficient, is more susceptible to fungal diseases than other avenacin-producing oat species (Osbourn et al., 1994). The increased susceptibility of chemically generated avenacin-deficient mutants of the diploid species *A. strigosa* to soil borne pathogens including the causal agent of take-all disease, *Gaeumannomyces graminis* var. tritici, provides further evidence for a role for avenacins in plant defence (Osbourn et al., 1994; Papadopoulou et al., 1999).

Avenacins have also been shown to lyse the zoospores of the oomycete root-infecting pathogens *Phytophthora cinnamomi* and *Saprolegnia litoralis* (Deacon and Mitchell, 1985). The antimicrobial activity of avenacins has been attributed to their ability to disrupt cell membranes via pore formation through complexing with sterols in the lipid bilayer (Augustin et al., 2011).

1.2.1. The structures of the four avenacins



Avenacin	R_1	R_2
A-1	NCH_3	OH
A-2	H	OH
B-1	NCH_3	H
B-2	H	H

Figure 1.1: Structures of oat root avenacins. Avenacins A-1, A-2, B-1 and B-2 are modified in different ways at the C-21 and C-24 atoms.

There are four different forms of avenacins: A-1, B-1, A-2 and B-2 (Burkhardt et al., 1964). Avenacin A-1 is the major avenacin found in oat roots (Crombie and Crombie, 1986). Avenacins (Figure1.1) consist of a β-amyrin backbone modified by hydroxylation, epoxyidation, glycosylation, and acylation (Crombie and Crombie, 1986). Avenacin A-1 and B-1 are acylated with *N*-methyl anthranilate, which gives these molecules strong bright-blue fluorescence, whereas avenacins A-2 and B-2 are modified with benzoate at the C-21 position and are not UV-fluorescent. Avenacin A-1 and A-2 are hydroxylated at the C-24 position while the avenacins B-1 and B-2 are not (Crombie and Crombie, 1986).

1.2.2. The avenacin biosynthetic pathway

Research on dissecting the avenacin biosynthetic pathway has been mainly carried out in the diploid oat species *A. strigosa*. Investigations of a collection of sodium azide-generated saponin-deficient (*sad*) mutants of *A. strigosa* defined at least seven genetic loci required for avenacin biosynthesis (Papadopoulou et al., 1999). Of these, so far five of the corresponding gene products have been cloned and characterized (Haralampidis et al., 2001; Mugford et al., 2013, 2009; Owatworakit et al., 2013; Qi et al., 2004, 2006). The first committed step (Figure1.2) of the avenacin pathway is the cyclisation of 2,3-oxidosqualene to β-amyrin, which is catalysed by the oxidosqualene cyclase (OSC) β-amyrin synthase bAS1 (SAD1) (Haralampidis et al., 2001). The β-amyrin backbone is

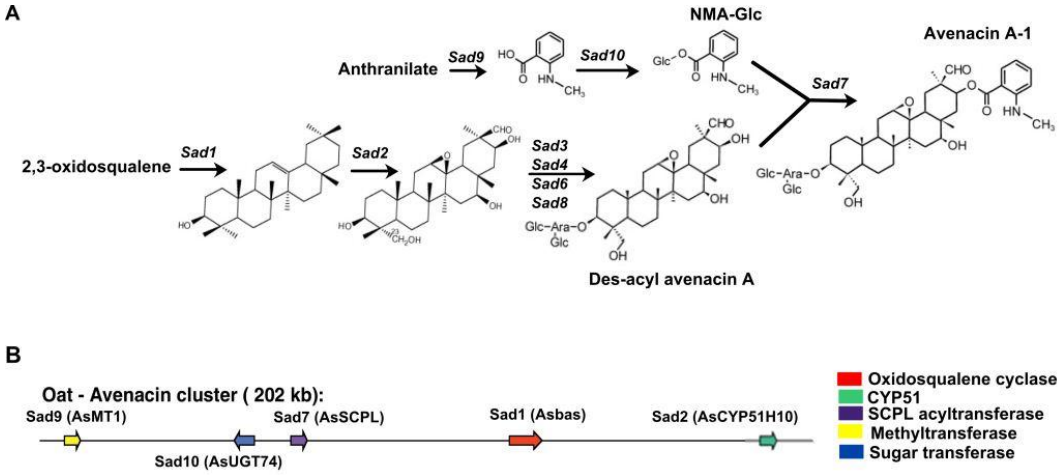


Figure 1.2: The avenacin biosynthetic pathway and the gene cluster. The genes are coloured according to the types of enzyme that they encode (see key).

then modified by the cytochrome P450 CYP51H10 (SAD2) by C12-13 epoxidation and C16 hydroxylation (Geisler et al., 2013). The triterpene scaffold is further modified by a series of oxidation and glycosylation steps to give des-acyl avenacin A (DAA) (Mugford et al., 2013; Mylona et al., 2008). In parallel, anthranilate from the shikimate pathway is modified by the methyl-transferase MT1 (SAD9) to give *N*-methyl anthranilate (NMA), which is the preferred substrate for the glucosyl transferase UGT74H5 (SAD10) (Mugford et al., 2013; Owatworakit et al., 2013). It was found that UGT74H5 (SAD10) and its homologue UGT74H6 are responsible for the synthesis of the acyl donors *N*-methylanthraniloyl-β-D-glucopyranose (NMA-glc) (the acyl donor for synthesis of avenacins A-1 and B-1) and benzoyl-β-D-glucopyranose (for avenacins A-2 and avenacin B-2), respectively (Mugford et al., 2009; Owatworakit et al., 2013). Finally, the acyl-transferase SCPL1 (SAD7) utilises these acyl glucose donors to transfer NMA-glc or benzoate to the des-acyl avenacin to generate avenacin A-1 and B-1 or A-2 and B-2 (Mugford et al., 2009).

1.2.3. Subcellular organisation of avenacin biosynthesis in the root epidermal cells

The early pathway enzymes SAD1 and SAD2 are likely to be associated with the endoplasmic reticulum (ER), most likely the smooth endoplasmic reticulum that is involved in steroid metabolism (Wegel et al., 2009) (Figure1.3). Further hydroxylation and oxidation steps then take place at the ER or in the cytosol to generate DAA (Mugford et al., 2013). Immunogold labelling using antisera raised against MT1, UGT74H5 and SCPL1 revealed SAD9 and SAD10 to be cytosolic, while SAD7 localised to the vacuole (Mugford et al., 2013). Thus, it

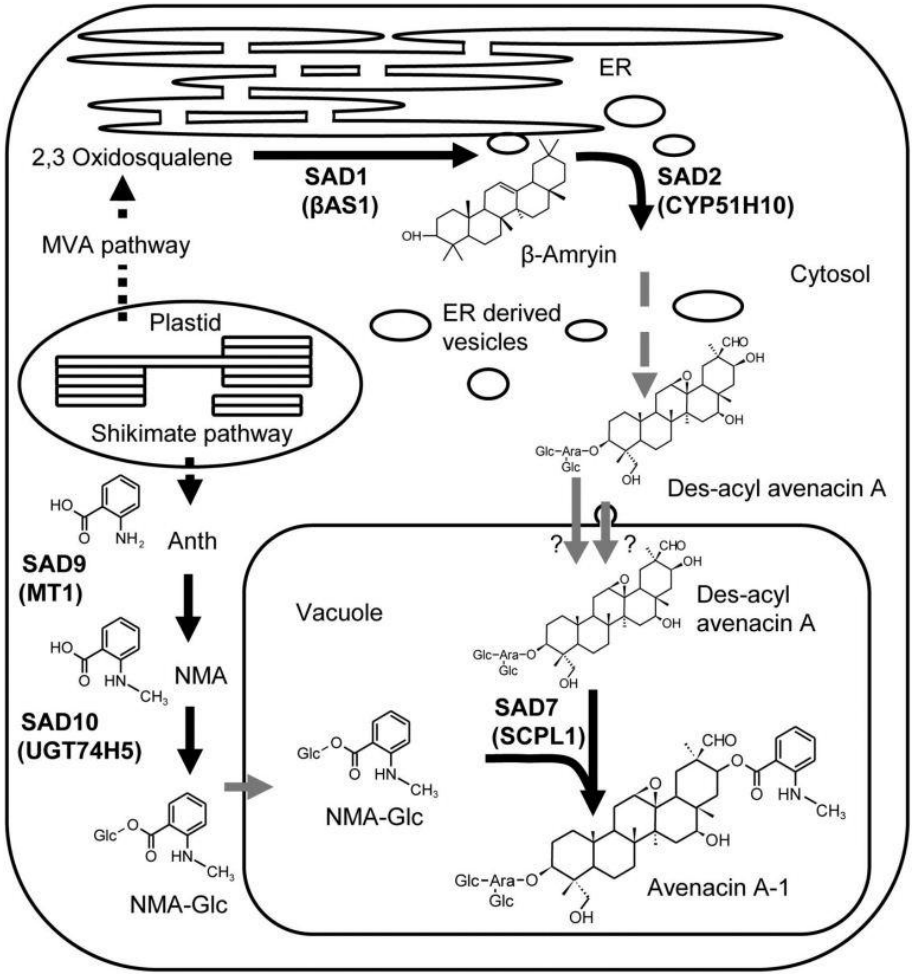


Figure 1.3: Subcellular organisation of avenacin biosynthesis. The pathway for the synthesis of avenacin A-1 is shown. The two precursors of avenacin A-1, 2,3-oxidosqualene and anthranilate (Anth) originate from the mevalonate and shikimate pathways, respectively. 2,3-Oxidosqualene is converted to β-amyrin by SAD1 (βAS1) and then further modified by SAD2 (CYP51H10) at the endoplasmic reticulum (ER). The triterpene scaffold is then further modified by a series of uncharacterized oxidation and glycosylation steps to give des-acyl avenacin A (DAA). Meanwhile, anthranilate is modified by SAD9 (MT1) and SAD10 (UGT74H5) to give *N*-methyl anthraniloyl-O-glucose (NMA-Glc), the activated acyl donor required by the acyltransferase SAD7. The triterpene glycoside and NMA-Glu are transported into the vacuole where NMA is transferred to the C-21 position of the triterpene glycoside by SAD7 (SCPL1) to give avenacin A-1. Reproduced from Mugford et al. (2013).

has been proposed that anthranilate derived from the plastid-localised shikimate pathway is modified by SAD9 in the cytoplasm to give *N*-methyl anthranilate (NMA) (Mugford et al., 2013). NMA is then glucosylated by SAD10 also in the cytosol, to give NMA-Glc (Mugford et al., 2013). DAA and NMA-Glc are transported to the vacuole (by as yet unidentified mechanisms) where the NMA group is transferred onto the C-21 position of DAA by SAD7 (Mugford et al., 2013, 2009).

Sad1, 2, 7, 9 and *10* all display root-tip specific expression (Haralampidis et al., 2001; Mugford et al., 2013, 2009; Owatworakit et al., 2013; Qi et al., 2006). Pathway-specific transcription factors for avenacin biosynthesis have not yet been identified. Fluorescence *in situ* mRNA hybridisation experiments have shown that the region of the chromosome encompassing the *Sad1* and *Sad2* genes is more decondensed in the oat root epidermal cells than in cells where the pathway is not active, suggesting that the avenacin gene cluster is also subject to regulation at the level of chromatin remodelling (Wegel et al., 2009). Importantly, these mRNA *in situ* hybridisation experiments also suggest that the sterol and avenacin pathways are inversely regulated at the level of transcription and are spatially separated, avenacin biosynthesis occuring in the root tip and sterol biosynthesis in the older part of the root (Wegel et al., 2009). The inverse spatial expression pattern of the *Sad* genes and sterol biosynthesis genes suggests that there may be competition between the two pathways, possibly due to substrate competition or to detrimental effects of avenacin production within cells that actively produce sterols (Wegel et al., 2009).

1.3. The avenacin gene cluster

Investigations into the genomic locations of *Sad1, 2, 7, 9*, and *10* revealed that these five avenacin biosynthetic genes are co-localised within a 202 kb subtelomeric region of linkage group AswC in *Avena strigosa* accession S75 (Figure 1.2b) (Qi et al., 2004). *Sad3, 6* and *8* are linked to *Sad1* but have not yet been cloned (Mylona et al., 2008; Papadopoulou et al., 1999; Qi et al., 2004). *Sad1, 2, 7*, and *9* are orientated in the same direction and are spaced at approximately 60 kb intervals (Qi et al., 2004). *Sad10* is inversely orientated relative to the other *Sad* genes, lying between *Sad9* and *Sad7*. The region that the avenacin gene cluster resides in does not share synteny with the rice genome (Qi et al., 2004), suggesting either this subtelomeric region has been newly formed in oats or that it was highly dynamic and has undergone extensive rearrangement since the divergence of oats from rice.

1.4. Gene clusters – an emerging field in plant biology

Neighbouring genes that are functionally related (co-functional genes) prevail in many organisms, especially in bacteria, archaea, fungi, and lower eukaryotes (Hurst et al., 2004). The most common examples of clustered genes are bacterial operons. So far 20 operon-like gene clusters for specialised metabolic pathways have been discovered in plants including the avenacin gene cluster (Itkin et al., 2013; Kliebenstein and Osbourn, 2012; Krokida et al., 2013; Matsuba et al., 2013; Winzer et al., 2012). These gene clusters consist of functionally-related genes that have not originated simply by tandem duplications. They consist for the most part of non-homologous genes that encode different classes of biosynthetic enzymes. Some examples of gene clusters for the synthesis of specialised metabolites in other plant species are described below.

1.4.1. The thalianol and marneral gene clusters

Two gene clusters for the synthesis of triterpenes have been characterised in *Arabidopsis thaliana*, the thalianol and the marneral clusters (Figure1.4) (Field et al., 2011; Field and Osbourn, 2008). The thalianol gene cluster (Figure1.4b) consists of four co-expressed genes: *At5g48010* encoding the thalianol synthase (THAS), *At5g48000* encoding the cytochrome P450 thalianol hydroxylase (CYP708A2/THAH), *At5g47990* encoding the thalian-diol desaturase (CYP705A5/THAD) and *At5g47980* encoding an acyltransferase (ACT) (Field and Osbourn, 2008). The marneral gene cluster (Figure 1.4c, d) consists of three coexpressed genes: *At5g42600* encoding the marneral synthase (MRN1), *At5g42590* encoding the marneral oxidase CYP71A16 (MRO) and *At5g42580* encoding a desaturase CYP705A12 (Field et al., 2011). THAS and MRN1 are both OSCs and catalyse the first steps in thalianol and marneral biosynthesis respectively, operating as branchpoint enzymes that divert the primary sterol synthesis precursor 2,3-oxidosqualene to triterpene synthesis (Field et al., 2011; Field and Osbourn, 2008). THAH, THAD and ACT are tailoring enzymes of the thalianol pathway while MRO and CYP705A12 are tailoring enzymes of the marneral pathways (Field et al., 2011; Field and Osbourn, 2008).
Importantly, the *OSC/CYP705* gene pairs of the thalianol and marneral gene clusters were likely to be derived from an ancestral *OSC/CYP705* gene pair during the expansion of *OSCs* and *CYPs*, after the *Brassicaceae* α whole genome duplication event (Field et al., 2011). The different orientation of the *OSC/CYP705* gene pair in the two gene clusters and the independent

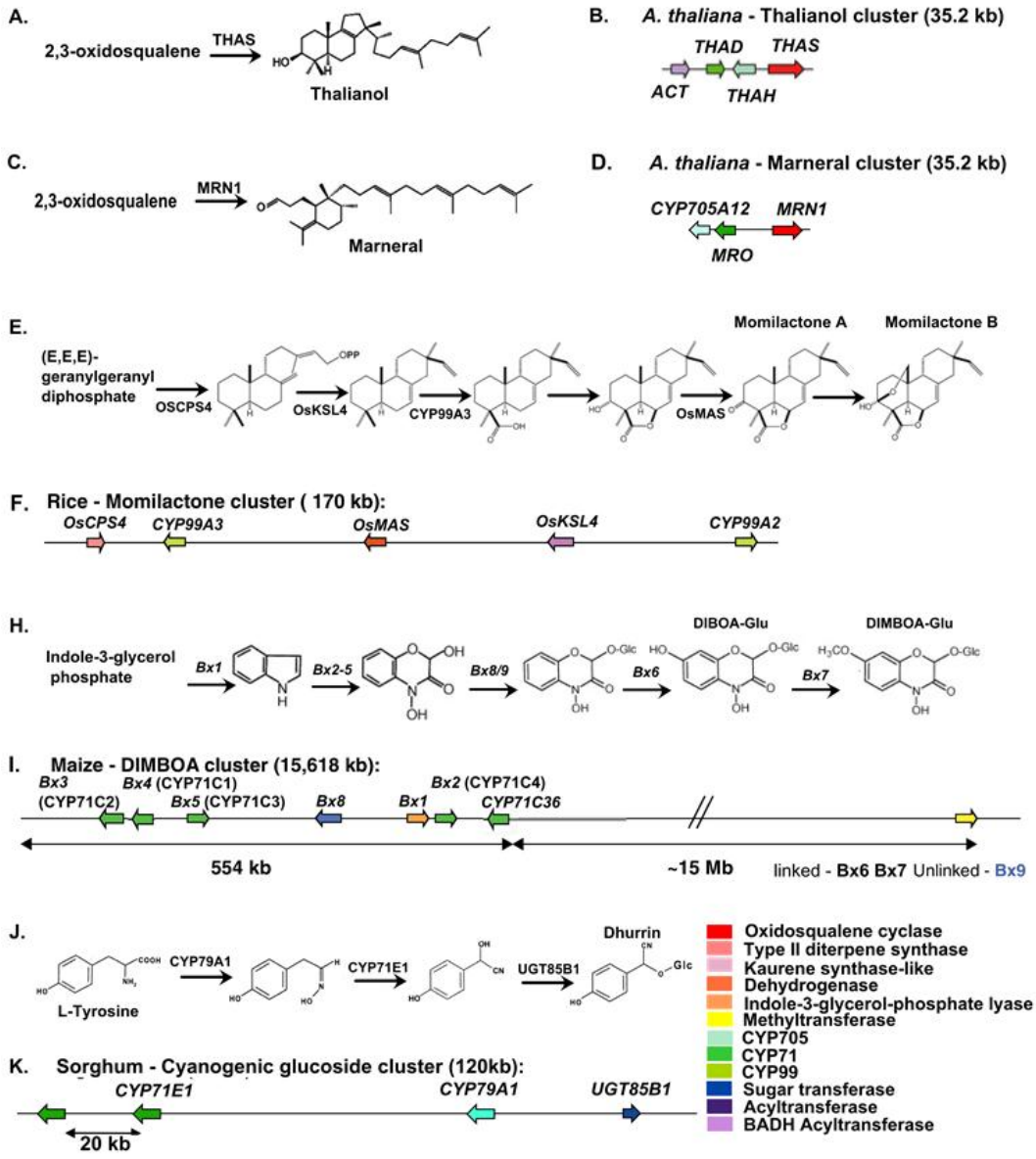


Figure 1.4: Five examples of specialised metabolic gene clusters. Schematic diagram of the pathway and the gene cluster structure of the corresponding clustered genes for A + B) thalianol (Field and Osbourn, 2008), C + D) marneral (Field et al., 2011), E + F) momilactones (Swaminathan et al., 2009), H + I) DIBOA and DIMBOA (Frey et al., 2009) and J + K) dhurrin (Takos et al., 2011). Intervening genes in the gene clusters that are not reported to encode gene products contributing to the specialised pathway are not shown. The genes are coloured according to gene function and enzyme family. The chemical structures of the pathway intermediates and products are shown. Adapted from Chu et al. (2011).

recruitment of genes for downstream tailoring enzymes to each cluster suggests that independent evolution gave rise to the thalianol and the marneral gene clusters in *A. thaliana* (Field et al., 2011). Both gene clusters are located in dynamic genomic regions that are enriched in transposable elements, which may have facilitated cluster assembly through recruitment of new genes to the region from gene families prone to ectopic transposition via gene relocation (Field et al., 2011).

1.4.2. The momilactone gene cluster

Diterpene biosynthetic genes have been observed to be in physical proximity in rice and tomato, suggesting the predisposition of these genes to clustering. The momilactone gene cluster in *O. sativa* is reviewed here as an example of a diterpene gene cluster (Figure 1.4e, f). Momilactone B has been shown to play a critical role in allelopathy and its release increases under low-nutrient levels or in the presence of weeds (Kato-Noguchi, 2009, 2011; Kato-Noguchi et al., 2010).
The momilactone gene cluster is situated in a region of 170 kb in the pericentric region of the p arm of rice chromosome 4 and contains five genes, *LOC_Os04g09900* encoding a *syn*-CPP synthase (Os-CPS4, a class-II diterpene synthase), *LOC_Os04g10060* encoding a kaurene synthase-like *syn*-pimaradiene synthase (Os-KSL4, a class-I diterpene synthase), *LOC_Os04g10010* encoding a dehydrogenase (Os-MAS1) that catalyses the final step in momilactone A biosynthesis, and two closely related cytochrome P450s genes (*LOC_Os04g10160* encoding a CYP99A2 that is likely to be non-functional, and *LOC_Os04g09920* encoding a multifunctional diterpene oxidase CYP99A3) (Figure 1.4e) (Swaminathan et al., 2009; Wang et al., 2011). In momilactone biosynthesis, Os-CPS4 converts the universal diterpenoid precursor (E, E, E,)-geranylgeranyl diphosphate (GGPP) to *syn*-copalyl diphosphate (*syn*-CPP) followed by downstream cyclisation to *syn*-pimera-7,15-diene by Os-KSL4. All five genes in the cluster are activated by OsTGAP1, a chitin oligo-saccharide elicitor-inducible basic leucine zipper transcription factor (Okada et al., 2009) that binds to upstream W-box elements of the genes (Nemoto et al., 2007; Okada et al., 2009). Phylogenetic studies suggest that the momilactone biosynthetic pathway genes Os-CPS4, CYP99A3, and Os-KSL4 originate from the gibberellic acid (GA) biosynthetic pathway (Chu et al., 2011; Swaminathan et al., 2009). A gene duplication event followed by neo-functionalisation of the core modular enzymes Os-CPS1, Os-KSL1, and CYP71 then led to the evolution of novel momilactone biosynthetic functions (Swaminathan et al., 2009). The multi-functionality of KS-L enzymes (Morrone et al., 2011) and of the CYPs (Wang et al., 2011)

invovled in rice diterpene synthesis are likely to have facilitated the transformation from gibberellin to momilactone biosynthesis. Although phylogenetic analyses appear to suggest that simultaneous gene clustering took place after the recruitment event of Os-CPS1/Os-KSL1 gene pair (Swaminathan et al., 2009), whether this gene clustering event led to neo-functionalisation for diterpene biosynthesis or the other way round was not further investigated.

1.4.3. The benzoxazinoid biosynthetic gene cluster

The ability to produce 2,4-dihydroxy-2*H*-1, 4-benzoxazin-3(*4H*)-one (DIBOA) and its deriative DIMBOA is found in many monocot species including maize, and also in isolated species within three dicot families (Frey et al., 2009; Grün et al., 2005; Schullehner et al., 2008). Here, the evolution of DI(M)BOA biosynthesis in both monocots and dicots is discussed as an example of repeated evolution (in dicots), and lineage specific loss, split and maintenance of the gene cluster (in monocots) (Dick et al., 2012; Frey et al., 2009; Grün et al., 2005; Nomura et al., 2003; Schullehner et al., 2008; Sue et al., 2011).

The first committed step in DIBOA biosynthesis is the conversion of indole-3-glycerolphosphate to indole (Figure 1.4h) in the plastid by the branchpoint enzyme BX1 (an indole synthase), which diverts precursors for synthesis of tryptophan to bezoxazinoid (Dick et al., 2012). Indole is then modified by a set of four closely related CYP71Cs (BX2-BX5) to DIBOA in the microsomes/ER. Subsequent steps involve glucosylation, carried out by BX8/BX9 and further downstream modifications by BX6 and BX7 in the cytosol to yield DIMBOA-Glu (Frey et al., 2009).

Bx1-8 form a gene cluster in the subtelomeric region of *Zea mays* chromosome 4 (Figure 1.4i) (Frey et al., 2009). Phylogenetic studies suggest that the DIBOA pathway arose from the duplication and neofunctionalisation of the tryptophan synthase alpha (*Tsa*) gene and recruitment of the set of four *CYP71Cs* prior to the radiation of *Poaceae* (Dutartre et al., 2012; Frey et al., 2009; Sue et al., 2011). Positive selection occurred during the evolution of *Bx2-Bx5*, leading to their distinct substrate specificities (Dutartre et al., 2012). In genomic studies carried out in monocots, it was concluded that physical clustering of the *Bx1* and *Bx2* ancestral sequences prompted co-evolution of both genes and initiated the formation of the DIBOA cluster in the *Poaceae* ancestor (Dutartre et al., 2012; Sue et al., 2011).

The gene cluster was later lost repeatedly in several barley accessions, *O. sativa, B. distachyon* and *S. bicolor*, whereas the DIBOA gene cluster was split into two chromosomal regions in wheat and rye (Grün et al., 2005; Nomura et al., 2003;

Sue et al., 2011). Of note, the splitting of the gene cluster in wheat and rye does not appear to have compromised the co-expression patterns of the *Bx* genes in young seedlings (Sue et al., 2011). In barley, synthesis of DIBOA and another specialised metabolite gramine in different accessions is mutually exclusive,which is presumed to be due to substrate competition for indole 3-glycerol phosphate. Clustering of *Bx* genes may be advantageous because the *Bx* genes could be co-acquired or co-deleted within the genome according to the presence/absence of antagonistic biosynthetic pathways (Grün et al., 2005).
Although Bx cluster is an ancestral genomic structure in the grasses, the branch-point enzyme for DIBOA synthesis BX1 and the key activation enzyme DIBOA-specific glucosyltransferase (BX8) have been reported to have evolved repeatedly in the Ranunculales and Lamiales by the neo-functionalisation of the *Tsa* gene (Dick et al., 2012; Schullehner et al., 2008). These repeated evolutions may be achieved by frequent duplication of indole synthases (IGL) and substrate promiscuity of UGTs (Dick et al., 2012; Dutartre et al., 2012; Yonekura-Sakakibara and Hanada, 2011).

1.4.4. Cyanogenic glucoside gene clusters

Biosynthesis of cyanogenic glucosides has been identified in both plants and arthropods for defense (Jensen et al., 2011; Takos et al., 2011). Although the end products are structurally diverse, cyanogenesis pathways in different plant lineages have evolved through repeated recruitment of the same classes of cytochromes P450 (CYP79s) and sugar transferases (UGT85s) (Takos et al., 2011). Here, the cyanogenic glycoside gene clusters are discussed as an example of repeated gene cluster formation.
The first committed step in cyanogenesis involves the branchpoint enzyme CYP79A1 which converts L-tyrosine (exclusively in *S. bicolor*), L-valine and L-isoleucine to oxime intermediates (Figure 1.4j) (Bak et al., 2006). The oxime is modified by NADPH-dependent dehydration and a C-hydroxylation by CYP71E1 followed by glycosylation by UGT85B1 to give cyanogenic glycosides (Bak et al., 2006; Jensen et al., 2011; Takos et al., 2011). In sorghum, the two CYP71s and the UGT form a membrane-bound complex for dhurrin biosynthesis, in order to facilitate channelling of toxic and unstable intermediates (Bak et al., 2006). Expression of dhurrin biosynthesis genes is highest during germination and early seedling development and then gradually decreases, but is induced by nitrate in older plants (Busk and Møller, 2002).
Cyanogenic glycoside biosynthetic gene clusters have been found in *Lotus japonicas, Sorghum bicolor* (Figure1.4k) and *Manihot esculenta* (Takos et al.,

2011). These three gene clusters all contain CYP79, CYP71E, and UGT85 genes that have evolved via species- or lineage-specific gene duplications. However the clusters do not share any similarity in terms of gene cluster structure, gene orthology, or synteny, suggesting that these clusters are examples of repeated evolution (Takos et al., 2011). In contrast, repeated evolution of cyanogenesis in the caterpillar *Zygaena* was achieved by recruitment of different classes of P450 and UGT genes compared to those responsible for cyanogenic glucoside biosynthesis in plants (Jensen et al., 2011).

1.5. Similarities among plant gene clusters

By comparing and contrasting the features of the plant metabolic gene clusters that have been characterised so far we may be able to gain insights into the likely functional significance of clustering and the mechanisms of cluster assembly. Firstly, the maize DIBOA cluster and the oat avenacin gene clusters are both located in the subtelomeric regions of their respective chromosome (Frey et al., 2009; Qi et al., 2004), regions that are prone to genomic rearrangement and so may facilitate cluster formation (Eichler and Sankoff, 2003). The thalianol and marneral gene clusters are also located in transposable element (TE)-rich dynamic chromosomal regions (Field et al., 2011). Secondly, most of the genes in the oat avenacin gene cluster and the rice momilactone gene cluster are spaced equidistantly (Mugford et al., 2013; Qi et al., 2004; Swaminathan et al., 2009). Thirdly, the first committed enzymes in the pathways for the plant gene clusters described so far have all arisen (directly or indirectly) by gene duplication from primary metabolic enzymes (Chu et al., 2011). Furthermore, the genes within these gene clusters are co-regulated via complex mechanisms, allowing the pathways to be differentially expressed at particular developmental growth stages and/or in response to external stresses (Busk and Møller, 2002; Dick et al., 2012; Okada et al., 2009; Wegel et al., 2009). Finally, tandem copies of closely related *CYPs* that carry out different pathways steps can be found in the maize DIMBOA (Frey et al., 2009) and the sorghum dhurrin gene clusters (Takos et al., 2011), suggesting that further duplication and neo-functionalisation of clustered genes may provide another dimension for pathway evolution, so generating more complex metabolites.

1.6. Gene clusters across kingdoms

Gene clusters are common features of prokaryotic and lower eukaryotic genomes (Koonin, 2009). Nonetheless, examples of gene clusters and even operons have

been identified in higher organisms such as the worm *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* (Nannapaneni et al., 2013). The best-known gene clusters in animals are the Hox gene clusters that determine the body plan in early embryogenesis (Lemons and McGinnis, 2006). As in plants, microbial gene clusters encode secondary metabolites such as antibiotics and toxins (Fischbach et al., 2008; Osbourn and Field, 2009). These gene clusters/operons may be laterally transferred between micro-organisms by plasmids, enabling rapid environmental adaption (Cooper et al., 2010; Norris and Merieau, 2013). Because numerous operons and gene clusters in lower eukaryotes have been well characterized and the field of operon formation is well-established in prokaryotic models, several hypotheses of operon formation and maintenance will be discussed later in this Chapter since this may aid our understanding of gene cluster formation in higher eukaryotes.

1.7. Genetic innovation and metabolic diversification

Most gene cluster formation hypotheses assume that co-functionality of genes subjected to clustering has already been established prior to clustering (Al-Shahrour et al., 2010). However, very little is known about the events associated with formation of novel pathways. Because the formation of gene clusters for new specialised metabolic pathways is closely associated with the establishment of gene functional neighbourhoods, gaining knowledge of how new genes and pathways form is vital to aid our understanding of different models of gene cluster formation. Metabolic pathway innovation can be divided into two main areas of research: 1) formation of new genes and 2) formation of new pathways/metabolic modules (Fondi et al., 2009; Kliebenstein and Osbourn, 2012).

1.7.1. Evolution of novel enzymes

New genes can originate by multiple mechanisms, mainly classified into two main catergories, gene duplication and *de novo* formation (reviewed in Van de Peer et al. (2009); Wu and Zhang (2013)). Duplication of protein-coding sequences is the main source of raw material for gene innovations (Conant and Wolfe, 2008; Kersting et al., 2012). Following duplication, the duplicated sequences may be relocated in the genome or be laterally transferred to a foreign genome, resulting in gene fusion/fission (modular rearrangement), changes in regulation and or changes in functionality via mutations (Bornberg-Bauer and

Albà, 2013; Conant and Wolfe, 2008; Dagan et al., 2008; Innan and Kondrashov, 2010; Kersting et al., 2012; Wu and Zhang, 2013). Ectopic recombination between paralogues also gives rise to new genes (Christiaens et al., 2012).

De novo acquisition of functional genes involves the recruitment of non-protein coding DNA sequences (Carvunis et al., 2012; Wu and Zhang, 2013). Genes that have originated *de novo* were first thought to be rare but are now increasingly recognised as species-specific orphan genes present in sequenced genomes (Carvunis et al., 2012; Wu and Zhang, 2013). For example, genome-wide studies have revealed that 11% of *Drosophila melanogaster* genes have arisen *de novo* (Wu and Zhang, 2013). While genes that have originated *de novo* appear to account mainly for orphan functionalities that are restricted to particular species, duplicated genes are solely attributed to functional novelties that may be repeatedly evolved in distinct lineages (Kersting et al., 2012). Due to the fact that most novel secondary metabolic genes have evolved as a consequence of gene duplication events (Ober, 2010), the mechanisms and fates of duplicated genes will be discussed here.

The ability to retain duplicated sequences is key to eukaryotic life (Koonin, 2009). Duplicated sequences in prokaryotes are likely to be lost due to energetic trade off and duplication is mainly contributed by horizontal gene transfer (Koonin, 2009; Treangen and Rocha, 2011). On the contrary, gene duplications in eukaryotes mainly involve endogenous sequences, with very few cases of horizontal gene transfer (Gilbert and Cordaux, 2013; Renner and Bellot, 2012). Duplication of genes can take place via whole genome duplication, segmental duplication, unequal crossing over, repair of staggered breaks, non-homologous end joining, transposition, retrotransposition, and insertion of reverse-transcribed cDNA into the genome. Whole genome duplication (WGD) has occurred frequently during the evolution of land plants and has been proposed to be an adaption strategy to drastic environmental changes (Van de Peer et al., 2009). Tandem duplications, on the other hand, give rise to young genes that can undergo accelerated evolution, so enabling adaption to new niches (Wang, 2013).

1.7.2. Gene duplication as a source of metabolic novelty

There are three possible fates of protein-coding gene duplicates according to Ohno′s model, 1) pseudogenisation (gene loss or non-functionalisation), 2) subfunctionalisation (the activity of the original gene becomes shared by the duplicates) and 3) neo-functionalisation (the duplicate gains a new function that is different from that of the original gene) (Ohno, 1970). Ohno proposed that after gene duplication one copy would be subjected to purifying selection to

preserve its ancestral function, while the other copy would be under relaxed selective constraints and free to evolve a new function (neo-functionalisation) (Innan and Kondrashov, 2010; Ohno, 1970). Because gene duplication and deletion events occur in an stochastic manner while gene loss and retention are non-random due to selection, Ohno′s model fails to describe the immediate consequences of gene duplication. Pseudogenisation, subfunctionalisation and neo-functionalisation of genes is only achieved via relatively long-term selection for gene retention (Sikosek et al., 2012). Therefore, several recent models and hypotheses (Table 1.1) focusing on adaption before and immediately after gene duplications have been proposed and validated experimentally to augment Ohno′s gene duplication framework (Birchler and Veitia, 2012; Conant and Wolfe, 2008; De Smet et al., 2013; Innan and Kondrashov, 2010; Lynch and Force, 2000; Sikosek et al., 2012; Xue et al., 2010).

The dosage balance hypothesis and the dominant-negative mutation hypothesis (Table 1.1) gave explanations for preferential gene retention and gene loss after WGD (Birchler and Veitia, 2012; De Smet et al., 2013). Together with gene conversion with the original copy that prevents gene duplicates from non-functionalisation, the innovation, amplification and divergence (IAD) model predicts how tandem arrays are expanded and maintained (Conant and Wolfe, 2008; Xue et al., 2010). The duplication, degeneration, complementation (DDC) and the escape from adaptive conflict (EAC) models (Table 1.1) both describe sub-functionalization of gene duplicates; the DDC model implies that both duplication and subfunctionalization were selectively neutral whereas the EAC model assumes slight selective advantages of duplication and positive selection exerted on both the original and duplicated copies of the gene (Sikosek et al., 2012; Zhu et al., 2012).

Reconstruction of the evolutionary trajectories of the duplicates of genes for bi-functional enzymes in bacteria and yeast showed mutations improving both activities as well as mutations improving one activity at the expense of the other, so favouring both the IAD and the EAC models over the DDC model (Table 1.1) (Näsvall et al., 2012; Voordeckers et al., 2012). Another layer of complexity following gene duplication is concerned with epimutations (Klironomos et al., 2013). The high rate of epimutation may lead to simultaneous changes in the expression profile of duplicated genes and further to gene neo- or sub-functionalisation (Klironomos et al., 2013). A recent study of gene methylation in rice showed that segmental duplicates and tandem duplicates have distinct methylation patterns, leading to different modes of gene sub-functionalisation via differential gene expression (Jiang et al., 2013). Changes in expression levels following duplication have played a significant role

Model	Definition	Reference
The dosage balance hypothesis	Stoichiometric balance is important for interacting genes, especially for those encoding interacting proteins. Hence, interacting genes (dosage sensitive) tend to be duplicated and retained collectively in whole genome duplication (WGD).	Birchler and Veitia (2012)
Dominant-negative mutation hypothesis	Duplication of genes with high connectivity in a metabolic network provides additional targets for deleterious dominant mutations and thus selection favours these genes to remain as singletons.	De Smet et al. (2013)
Innovation, Amplification and Divergence (IAD) model	The ancillary function of a gene is favoured by dosage selection, resulting in gene duplications. New functions are evolved via accumulation of adaptive mutations.	Conant and Wolfe (2008)
Neo-functionalisation via originalization	Recombinations between gene duplicates restore the ancestral gene function of paralogues, resulting in prolonged time to non-functionalisation.	Xue et al. (2010)
Duplication, Degeneration, Complementation (DDC) model	Neutral mutations occur in all gene duplicates leading to sub-functionalisation of genes, in expression patterns and protein domains, such that the complementation of duplicated copies restores the total activities of the original gene.	Innan and Kondrashov (2010); Lynch and Force (2000)
Escape from Adaptive Conflict (EAC) model	Multifunctionalities of enzymes are selected for metabolic homeostasis and rapid adaption. Gene duplication allows specialisation of duplicates from the ancestral promiscuous enzyme for functional optimisation.	Sikosek et al. (2012)

Table 1.1: Evolutionary models of the fates of gene duplicates.

in gene retention and neo-functionalisation (Cheng et al., 2012; Jiang et al., 2013; Rodgers-Melnick et al., 2012; Schnable et al., 2012; Yang and dos Reis, 2011). However, gene originalisation may counteract duplicated gene non-functionalization and may prolong the time of gene preservation for neo-functionalisation to take place (Xue et al., 2010). Tandem duplicates and genes arising from WGD or segmental duplications also have distinct fates depending on the mode of duplication (Freeling, 2009; Gout et al., 2010; Jiang et al., 2013). In addition to the mode of duplication, the evolution of gene duplicates is mainly governed by gene functionalities and network interactions (Gout et al., 2010; Wu and Qi, 2010).

Gene 'duplicability' also correlates with the complexity of the organism (Zhu et al., 2012). WGD generates complete redundancy of metabolic networks and thus allows collective adaption of duplicated genes involved in the same existing pathway (De Smet et al., 2013). On the contrary, tandem duplications that increase the dosage of hubs or bottleneck enzymes in the metabolic network may have negative effects on flux sensitive pathways and protein-protein complexes requiring strict stoichiometric subunit proportions (Wu and Qi, 2010). On the other hand, a high rate of nucleotide substitution brought about by unequal recombination and increased mutational targets, accelerates sequence divergence among tandem arrays and leads to gene neo-functionalisation (Jiang et al., 2013; Wang, 2013).

An alternative way to create new genes (neo-functionalisation) from the wealth of duplicated genetic material from WGD is modular rearrangement of protein-coding genes (Kersting et al., 2012).

1.7.3. Origin of new metabolic pathways

When considering how new metabolic pathways are formed and how the new pathway is eventually incorporated into the genome in the form of new genes and new regulatory elements, it is important to consider the theme of metabolic networks. The term metabolic network refers to collections of cellular biochemical reactions (Larhlimi et al., 2011). Metabolic networks contain a high level of redundancy, where the same chemical products can be produced via alternative routes (Wang and Zhang, 2009). High redundancy contributes to both metabolic robustness (internal and external perturbations) and mutational robustness (elimination or gain of enzyme-coding genes) (Rodrigues and Wagner, 2009; Wang and Zhang, 2009). Redundant pathways may be regarded as having accumulated from adaption to the fluctuating environment as a by-product rather than retained by selective advantages for adaptive backup (Wang and

Zhang, 2009). However, the majority of redundant reactions in *Escherichia coli* and *Saccharomyces cerevisiae* are found to be selectively maintained (Wang and Zhang, 2009). The selection for maintain metabolic redundancy echo with the criteria of the EAC and the IAD model for gene retention, supporting the hypotheses that enzyme multifunctionality is favoured by selection (Sikosek et al., 2012) and redundant gene duplicates may be retained in the genome because of their neutral or even slightly positive effect on the phenotype (Wang and Zhang, 2009), thus allowing time for gene diversification to take place.

Another characteristic of metabolic networks is their plasticity (Rodrigues and Wagner, 2009). It has been demonstrated that the same phenotype of the metabolic network could be genotypically (reactions underlying the metabolic network) diverse; and that genotypically identical metabolic networks may have very different phenotypes (Rodrigues and Wagner, 2009). The plasticity of metabolic networks provides access to an enormous space of neighbouring network genotypes and phenotypes for rapid adaption as well as novel phenotypes (Rodrigues and Wagner, 2009). Network plasticity channels the increased dose of chemicals brought about by gene or genome duplication through *de novo* instantaneous formation of new alternative pathways within the existing network genotype (Larhlimi et al., 2011). In addition, the increase in occurrence of a chemical reaction would lead to the appearance of more alternative pathways (Rodrigues and Wagner, 2009). Thus a highly connected node (widely used metabolite) in a network may lead to the appearance of multiple alternative pathways. In contrast, nodes with low connectivity (uncommon metabolites) may have very few alternative pathways. This has been indirectly validated by investigations in *A. thaliana* showing that secondary metabolic pathways tend to have functional compensation only through highly similar gene duplicates because of a lack of alternative pathways (Hanada et al., 2011).

The current proposed mechanisms of new metabolic pathway formation are 1) *de novo* stepwise recruitment, and 2) modular evolution (Fondi et al., 2009). Although stepwise recruitment may be difficult to achieve because of the requirement for significant evolutionary benefits of pathway intermediates (Kliebenstein and Osbourn, 2012), it is highly feasible given the plasticity and robustness of the metabolic network. Once a new enzyme activity is found and has led to production of a novel metabolite (the first pathway intermediate), which is likely to be selectively neutral even regarded as "toxic" because it can be converted to multiple nontoxic end products by simultaneously formed alternative pathways (Ravasz et al., 2002). It is also likely that the newly formed pathway relies mainly on ancillary enzyme activities and can only produce these novel intermediates and end products in low amounts, so reducing

the detrimental effects of the "toxic" intermediates. Improved production of the selectively advantageous novel end product may drive gene duplication-diversifications leading to a stable operation of the alternative pathways (Lobkovsky and Koonin, 2012; Sikosek et al., 2012).

1.8. Plasticity of genome organisation

Selective pressures act indirectly on genome architecture because adaptive phenotypes do not necessarily relate to changes in genotype (Rodrigues and Wagner, 2009). For example, phenotypic changes brought about by changes in gene expression resulting from genome rearrangements can be remedied by rapid rewiring of regulatory networks.

The eukaryotic genome has been described as being quasi-random, different regions of the genome being under different selective pressures with regard to GC content, gene density and gene order (Koonin and Wolf, 2010). The genome layouts of higher eukaryotic organisms (land plants and vertebrates) are subjected to frequent genomic rearrangements, which are then preserved by neutral population-genetic processes (Koonin and Wolf, 2010). While there are intense constraints on genome layout in prokaryotic species, higher eukaryotes have fewer evolutionary constraints on gene order and genome structure due to their relatively small population sizes and long generation times (Koonin, 2009; Koonin and Wolf, 2010). Furthermore, the large portion of non-coding DNA in the genome and complex spatial organisation of eukaryotic genomes obscure the presence of functional gene neighbourhoods akin to those that can be readily identified in simple microbial genomes. Nonetheless, comparative genomic studies have provided compelling evidence that neighbouring genes in eukaryotic genomes are more likely to be functionally related and co-expressed (Al-Shahrour et al., 2010; Amoutzias and Van de Peer, 2008; Hurst et al., 2004). Because prokaryotes have large population sizes and high evolutionary selective pressure form streamlined genomes (Koonin and Wolf, 2010), most hypotheses and models about genome architecture evolution have been established for operon organisation. Therefore, concepts about the evolution of operons are discussed here with the aim of enhancing the understanding of eukaryotic gene cluster evolution.

1.9. Operons – the best known gene cluster model

Many of the genes in prokaryotic genomes are organised in structures known as operons (Képès et al., 2012). Operons consist of groups of co-localised genes that are transcribed under the control of a single operator into a polycistronic mRNA (Price et al., 2006). Genes in the same operon tend to be functionally related and co-regulated. It has been shown that *E. coli* genes encoding functionally related metabolic enzymes have a higher tendency to be clustered in an operon than those encoding protein-protein complexes (Kovács et al., 2009). Transcriptional-translational coupling, lack of intracellular compartmentalisation and avoidance of conflicts between genome replication and transcription impose strong selective constraints on the organisation of operons (Képès et al., 2012). Based on these physiological features of bacteria and frequent horizontal gene transfer, many hypotheses and models explaining operon evolution have been proposed (Ballouz et al., 2010).

The two earliest proposed models for gene cluster formation are the Natal model and the Fisher model (Ballouz et al., 2010). The Natal model (Table 1.2) hypothesises that tandem gene duplication and divergence results in arrays of genes arranged in proximity (paralogous gene clusters) involved in related metabolic pathways (Ballouz et al., 2010). However, the lack of sequence similarity shared by genes in operons and functional gene clusters can not be explained by the Natal model. The Fisher model (Table 1.2) suggests a mechanism for cluster formation and maintenance in which selection favours co-segregation (close linkage) of beneficial combination of alleles of co-adapted genes (Fondi et al., 2009; Price et al., 2006). A further interpretation of Fisher′s theory was that linked genes could be gained or lost as functional cassettes. However, high genomic recombination rates (which were proved to be rare in bacteria and eukaryote-specific), are required for operon formation under the Fisher model (Ballouz et al., 2010). Furthermore, physical proximity of genes does not enhance gene co-evolution, which may be significant for achieving gene co-functionality (Cohen et al., 2012). The selfish operon model (Table 1.2) proposes that horizontal gene transfer is the main source of bacterial gene clusters (Lawrence and Roth, 1996). However, the selfish operon model only holds in environments with low recombination, and high transfer rates (Ballouz et al., 2010). Additionally, clustering of essential genes and the specific gene orientation in functional gene clusters can not be explained by selfish operon model (Fang et al., 2008; Lawrence and Roth, 1996; Lim et al., 2011). In conclusion, the selfish operon model does not provide the sole explanation for gene clustering but HGT may act as one of the indirect

driving forces of gene clustering (Treangen and Rocha, 2011).

The co-regulation model, which hypotheses that gene co-expression is the driver behind operon formation (Table 1.2), had become a more widely accepted (Price et al., 2005) because gene co-regulation enhances pathway modularity (Espinosa-Soto and Wagner, 2010). The co-regulation model is further elaborated by the protein immobility model and the transcription noise model (table 1.2) which state that optimized gene order in an operon is selected to enable the stoichiometric ratios of gene products to be controlled (Kovács et al., 2009; Ray and Igoshin, 2012). According to the transcription noise model, operon organisation is the best solution to ensure that the correct proportions of proteins are translated (Ray and Igoshin, 2012). When comparing the noise difference (fluctuations of pathway intermediates) of unclustered and clustered *lac* genes, it was demonstrated that the primary consequence of lac gene clustering was a reduction in fluctuation of intracellular metabolites (Ray and Igoshin, 2012). It is further hypothesized that gene clustering would be particularly beneficial for promiscuous enzymes, conflicting pathways and poorly expressed pathways to ensure efficient metabolic fluxes (Ray and Igoshin, 2012). It was speculated that metabolic pathways formed by multiple operons should contain break points at non-toxic intermediates, allowing novel pathways to form in the manner of modular rearrangements (Ray and Igoshin, 2012).

The protein immobilisation model (PIM) (table 1.2) suggests that a local metabolome is formed readily when functionally related gene products are co-translated, so reducing biochemical interference by pathway intermediates and increasing output of the pathway (Kovács et al., 2009). The PIM model predicts that the impact of gene order rearrangement would be most pronounced on distant genes within an operon (Kovács et al., 2009) because of the increasing stochasticity in protein translation with increasing intergenic distance. Indeed, gene order preservation increased when the physical distance separating the members of gene pairs increases (Kovács et al., 2009). Besides the mathematical modelling evidence to support the PIM model (Kovács et al., 2009), *In vivo* experiments also provided insights into selection on operon gene order (Lim et al., 2011). Four sets of artificial operons of fluorescent protein genes were constructued and the protein production of each gene was measured, after transformation of the artificial operons into *E.coli* (Lim et al., 2011). It was shown that the position of a gene within an operon and the operon length are directly correlated to the expression level of the gene (Lim et al., 2011). Genes located at the $5'$ end of an operon may have increased expression due to the increased time period for translation during transcription (Lim et al., 2011). Thus the gene order of an operon could be exploited for fine tuning of expression

patterns (Lim et al., 2011). A comparative analysis between *Mycobacterium leprae* and *Mycobacterium tuberculosis* has demonstrated such position-dependent functional importance of genes within operons (Muro et al., 2011). *M. leprae* is phylogenetically close to *M. tuberculosis* but 50% of its genome has undergone pseudogenisation (Muro et al., 2011). In the analysis, gene 'essentialness' (using pseudogenes in *M. leprae* as a marker of dispensability) and gene location between the two species were compared (Muro et al., 2011). The analysis showed that genes at the 5′ end of operons in *M. leprae* tend to be essential genes in both species; and that pseudogenes in *M. leprae* tend to be located at the 3′ end of operons and their orthologues tend to be non-essential genes in *M. tuberculosis* (Muro et al., 2011). This again showed that gene order in operons is under selection.

The gene persistence model (Table 1.2) offers an explaination for essential (persistent) gene clustering (Fang et al., 2008). Fang and coworkers hypothesized that clustering of highly persistent (essential) genes provides protection from gene non-functionalisation by reducing mutational targets (Fang et al., 2008). Although the persistent genes are arranged in operons, these persistent operons distribute evenly across the genome and thus do not potentially reduce the number of mutational targets (Bratlie et al., 2010). Analysis of persistent genes concluded that weak operon proteins (proteins with weak tendency for operon participation) shared more interacting partners, evolved more slowly, and tended to be longer than the strong operon proteins (proteins with strong tendency for operon participation) (Bratlie et al., 2010). It was suggested that because weak operon genes shared more interacting partners and thus were under more selective constraints leading to slow evolution, that weak operon genes may be involved in multiple pathways and thus were not favoured by specific transcriptional regulation with a restrictive set of interacting partners offered by operon organisation (Bratlie et al., 2010). This implies that gene clustering is dependent on gene function and network connectivity.

The scribbling pad model (Table 1.2) suggests that accessory genetic material provides a reservoir for genetic exchange, a 'scribbling pad' for operon formation (Table 1.2) (Norris and Merieau, 2013). The recombination rate is much higher in plasmids than in chromosomes (Norris and Merieau, 2013). Therefore operons with beneficial combinations of genes can be formed frequently by trial-and-error in plasmids and eventually integrate into the chromosome via recombination between plasmids and chromosomes (Cooper et al., 2010; Norris and Merieau, 2013). The scribbling pad model is strongly supported by recurring cluster and operon assembly and the accelerated gene evolution in accessory genetic regions (Cooper et al., 2010; Fischbach et al., 2008; Martin and McInerney, 2009).

Table 1.2: Models and hypotheses about operon evolution

Evolutionary model	Hypothesis of the model	Evidence supporting the model	Evidence against the model	Reference
The Natal model	Gene clusters arose by tandem gene duplication and diversification.	Paralogous gene clusters, such as the Hox cluster, are found in eukaryotes.	The model is unable to explain the formation of operons and functional gene clusters, which contain genes that do not share sequence similarity.	Ballouz et al. (2010); Fang et al. (2008)
The Fisher model	Selection confers advantages of tightened linkage (and thus reduced recombination) between genes with related function (eg. that are involved in the same pathway)	Close linkage of a pair of polymorphic genes in T-even phages are driven by protein-protein interactions	Essential genes or genes with similar expression patterns (genes with no functional relationship) form gene clusters. Genes in operons are frequently replaced via recombination.	Bratlie et al. (2010); Fang et al. (2008); Martin and McInerney (2009)
The selfish operon model	Genes aggregate into a cluster to enhance co-transfer into other organisms.	Rare gene clusters could be obtained by horizontal gene transfer (HGT)	Computational modelling of *E. coli* shows that genes in common operons do not always show sequence homologies.	Lawrence and Roth (1996); Pál and Hurst (2004)
The co-regulation model	Operon structure reduces the information required to specify the expression patterns for several co-regulated genes and so is selectively advantageous.	High conservation of regulatory sequences have been reported in *E. coli* and *Bacillus subtilis*.	There is no evidence to suggest that co-regulation can drive operon formation. Thus, co-regulation is likely to contribute to cluster maintainance but is not sufficient to drive cluster formation.	Espinosa-Soto and Wagner (2010); Hermsen et al. (2006); Lercher and Hurst (2006)

continued on next page

continued from previous page

The transcription noise hypothesis	Operon organisation reduces intrinsic noises in transcription, translation and metabolism.	Essential genes colocalise in open-chromatin regions in yeast.	Insufficient to explain gene clusters located in high noise domains such as secondary chromosomes.	Batada and Hurst (2007); Lim et al. (2011); Ray and Igoshin (2012)
The protein immobility model	Clustered genes are colinear to pathway reaction steps to minimize stalling of metabolism due to stochastic protein loss.	Lowly expressed operons contain more collinear gene orders.	Gene orders of metabolic operons are not highly conserved across bacterial species.	Kovács et al. (2009)
The gene persistence model	Selective pressures act on maintaining persistent genes in a genome that is subjected to frequent deletions, driving gene clustering.	Persistent genes are mainly located in operons in stable bacterial genomes.	Persistent operons are distributed across the genome.	Bratlie et al. (2010); Fang et al. (2008)
The scribbling pad model	Plasmids and integrative conjugative elements play a central role in operon assembly.	Operons are enriched in plasmids and gene order is divergent in conserved operons across species.	Restoration of dead operons by removal of intervening genes has not yet been observed.	Martin and McInerney (2009); Norris and Merieau (2013)

Overall, each of these hypotheses and models only partially describes this multifaceted evolutionary process. Nonetheless, concepts from some of the operon formation models are compatible with the eukaryotic system and may aid rationalisation of our understanding of the selective pressures exerted on eukaryotic gene clustering.

1.10. Functional neighbourhoods in eukaryotic genomes

Present-day eukaryotic genomes are the result of the interplay of genome rearrangements and network evolution to produce or maintain adaptive phenotypes (Grassi and Tramontano, 2011; Hurst et al., 2004; Koonin and Wolf, 2010; Rodrigues and Wagner, 2009). Plant genomes are especially dynamic due to increased genome rearrangements brought about by polyploidisation, invasions of transposable elements and post-WGD diploidisation (Murat et al., 2012, 2010; Wicker et al., 2010).

In a study of gene functional neighbourhoods, which refine to functionally related and co-localised genes, in eight eukaryotic model species, it has been discovered that up to 12% of non-homologous genes in mice are arranged in functional neighbourhoods (Al-Shahrour et al., 2010). For *A. thaliana*, 193 functional neighbourhoods have been reported (3% of genes in the genome) (Al-Shahrour et al., 2010). Interestingly, the functional neighbourhoods shared between humans and chimpanzees were significantly enriched in synteny breakpoints, with low-orthology neighbourhoods containing more synteny breaks and a lower degree of co-expression compared to high-orthology neighbourhoods (Al-Shahrour et al., 2010). This provided strong evidence that selection on functional neighbourhoods can be exerted on two different levels that favour: 1) continuous reorganisation or construction of functionally clustered genes and 2) preservation of optimized clusters (Al-Shahrour et al., 2010). Furthermore, it was concluded that selection on functional gene neighburhoods operates at the level of gene function but not on the genic level, thus leading to repeated construction of new functional neighbourhoods, via genome rearrangement and rewiring of gene expression, to replace lost neighbourhoods (Al-Shahrour et al., 2010).

These studies (Al-Shahrour et al., 2010) show that: 1) functionally related genes are selected to be clustered; 2) clustering involves active genomic rearrangements and; 3) well-established co-expression patterns are associated with cluster conservation. It is obvious that eukaryotic gene clusters and prokaryotic operons evolve in similar manners (repeated gene clustering) under similar selective pressures (selection on functional relatedness and co-regulation).

1.11. Mechanisms and selective constraints in eukaryotic gene cluster formation

Three interdependent elements key to gene cluster formation, 1) achieving gene physical proximities, 2) establishing specific functional relatedness and 3) optimizing gene product stoichiometry, can be summarised from all these studies so far. The complexities of eukaryote genomes requires broad consideration of a wide range of factors to rationalize the quasi-random genome layout in eukaryotic genomes (Koonin and Wolf, 2010). Firstly, eukaryotes possess large, expansive genomes the majority of which consists of non-coding DNA (Koonin and Wolf, 2010). Instead of selection for streamlined genomes, eukaryotic genomes are under relatively relaxed selective constraints and have tremendous genome plasticity (Koonin and Wolf, 2010). Secondly, eukaryotic genomes differ from those of prokaryotes due to the physiological complexities of intracellular compartmentalisation and multicellular organisation in eukaryotes (Koonin, 2009; Koonin and Wolf, 2010). Thirdly, the epigenetic system in eukaryotes is complex and heritable (Aceituno et al., 2008; Fedoroff, 2012; Iyer, 2012; Klironomos et al., 2013), acting as a buffer to dampen the effect of genomic changes on the fitness of the organism. Furthermore, many of the intrinsic sequences governing differential expression of genes, nucleosome packaging and spatial arrangement of genomic DNA are embedded in the non-coding part of eukaryotic genomes (Fedoroff, 2012; Irimia et al., 2013; Iyer, 2012).

1.11.1. Achieving gene physical proximity

As discussed before, selection acts indirectly on the eukaryotic genome layout (Koonin, 2009). Therefore, it is incorrect to assume that selection causes genomic changes leading to gene cluster formation. Instead, genome rearrangement is a stochastic process and any fitness enhancing functional neighbourhoods established must be maintained by selection. Although gene clusters facilitate the co-inheritance of beneficial combination of allelles (Osbourn, 2010a), the mechanism of cluster assembly is likely to be random. Nonetheless, increases in genome instability and enrichment in chromosomal rearrangements have been observed as immediate responses to nutrient limitation to increase gene mutation rate for a higher occurence of beneficial adaptive allelles (Gresham et al., 2008). Formation or disruption of physical linkage of genes via genome rearrangement is frequent and widespread in eukaryotic genomes, brought about mainly via duplications and invasion of transposable elements (reviewed in Eichler and Sankoff (2003)). The avenacin and DIBOA clusters are both located in subtelomeric

regions and the momilactone cluster is located in the pericentromeric region (Frey et al., 2009; Swaminathan et al., 2009; Wegel et al., 2009). Genome rearrangements are particularly prevalent in these regions of chromosomes, which are abundant in tandem repeats and transposable sequences (Eichler and Sankoff, 2003). Studies in *A. thaliana* and yeast have shown that subtelomeric regions contain recently duplicated fragments from other chromosomes and are enriched in species-specific transposons (Brown et al., 2010; Wang et al., 2010). Subtelomeric genes in yeast have a high turn-over rate, elevated copy number variation and expression divergence, leading to rapid neo-functionalisation (Brown et al., 2010). Thus, subtelomeric regions may act as a scribbling pad (Norris and Merieau, 2013) for evolutionary innovation, aided by rapid gene rearrangement and chromatin-dependent silencing, the latter influenced by the abundance of TE elements (Eichler and Sankoff, 2003). Transpositions mediate genome rearrangements by creating double strand breaks during excision and providing homologous sequences for illegitimate recombination (Fedoroff, 2012; Wijchers and de Laat, 2011). Indeed, intergenic regions of the thalianol, marneral and noscapine clusters are interspersed with TE elements (Field et al., 2011; Winzer et al., 2012), which are potentially involved in genome rearrangements.
Functional neighbourhoods have been observed in numerous sequenced genomes (Al-Shahrour et al., 2010; Amoutzias and Van de Peer, 2008; Hurst et al., 2004) but very few studies have addressed the selective pressures that drive their assemby. A recent investigation of genomic islands of divergence showed that genome rearrangement leads to active relocation of positively-selected locally-adaptive loci to give tight physical linkage (Yeaman, 2013).

1.11.2. Establishing co-functionality within gene neighbourhoods

A feature of gene clusters is their high modularity but low network connectivity (Espinosa-Soto and Wagner, 2010). This may be attributable to the fact that gene clusters are enriched for secondary metabolic genes (Hanada et al., 2011). Alternatively, high modularity isolates the pathway from the central network, providing confinement of toxic pathway intermediates and reducing conflicts and interference between the biochemical modules (Espinosa-Soto and Wagner, 2010). High modularity is attained by high substrate specificities of enzymes and coordinated expression of pathway genes (Espinosa-Soto and Wagner, 2010), which can be attained by their physical linkage (Dutartre et al., 2012; Wagner, 2008).
The enrichment of TE elements in gene clusters (Field et al., 2011) may also

facilitate turn-over of genes under selection for co-functionality and hence accelerate the gain of functional relatedness among physically clustered genes (Al-Shahrour et al., 2010). In addition, genes located distantly in the genome but brought to spatial proximity in the nucleus by transcriptional regulators may attain physical linkage during the process of non-homologous double-strand break (DSB) repair (Wijchers and de Laat, 2011).

1.11.3. Optimizing gene product stoichiometry

Selection for coordinated expression is important for gene cluster maintenance (Al-Shahrour et al., 2010; McGary et al., 2013; Price et al., 2005). Lowly expressed gene functional neighbourhoods and operons with disrupted expression are lost easily (Al-Shahrour et al., 2010; Muro et al., 2011; Price et al., 2006). Gene expression in eukaryotes is dependent on gene intron-exon structure, proximal and distal cis-elements, the location of the gene in the genome, the neighbouring genes, the chromatin structure and the spatial orientation in nuclear space (Brown et al., 2010; Buetti-Dinh et al., 2009; De and Babu, 2010; Ebisuya et al., 2008; Fedoroff, 2012; Irimia et al., 2013).

Gene co-expression is positively correlated to physical co-localisation (Ebisuya et al., 2008; Espinosa-Soto and Wagner, 2010). Genes are found to cluster according to their expression breadth and expression rate (Batada and Hurst, 2007). Co-expression of neighbouring genes is a consequence of the transcription ripple effect led by the opening of chromatin structure (Ebisuya et al., 2008). For example, in yeast the transcription ripple spreads to a 100 kb radius, leading to co-activation of neighbouring genes by a locally highly expressed gene (Ebisuya et al., 2008). On the other hand, cis-regulatory elements of a gene may be located in introns or in the intergenic regions between neighbouring genes and consequently, gene activation requires relaxation of the chromatin packaging of neighbouring genes so as to allow access of trans-acting regulators to these enhancers, leading to transcriptional de-repression of the neighbouring genes (Irimia et al., 2013).

The spatial clustering of transcription factor proteins inside particular regions of the nucleus, such as transciptional factories, also exerts selective constraints on gene functional neighbourhoods (Cook, 2010; Janga et al., 2008). Specific transcription factors tend to preferentially bind to target genes on the same chromosome compared to targets on distinct chromosomes (Janga et al., 2008). This phenomenon has been rationalized by the PIM Model (Kovács et al., 2009), which suggests lowly expressed transcription factors have to be colocalised with their target genes to reach high protein concentration locally for effective

transcription activation.

Furthermore, eukaryotic genomes are spatially organized in loops by chromatin remodelling and the transcription machinery (Cope et al., 2010; De and Babu, 2010; Feschotte, 2008). For example, the 'looping in' and 'looping out' of genes in the *HoxA* cluster regulates differential gene expression within the cluster (Fraser et al., 2009). Positioning of the loops in the nucleus is highly constrained by nucleosome architecture and the overall spatial organisation of the chromosomes (Cope et al., 2010; Iyer, 2012; Madan Babu et al., 2008). Therefore, physical clustering of co-functional genes may be advantageous for co-regulation simply by enabling small changes in local chromatin conformation rather than by pulling DNA loops from different chromosomes into spatial proximity (Wijchers and de Laat, 2011).

Furthermore, eukaryotic genomes are divided into chromatin domains that are either enriched in essential genes or in lowly expressed genes (Batada and Hurst, 2007). Such organisation is shaped by the selection against transcriptional noise (Batada and Hurst, 2007). Highly expressed (usually essential) genes are preferentially clustered in open chromatin domains of low nucleosome occupancy, defined as low-noise domains, for stable transcription while lowly expressed genes are enriched in high-noise regions such as the subtelomeric domains of the chromosome (Batada and Hurst, 2007). Subtelomic regions are generally enriched in TEs that are targeted by epigenetic silencing through dense heterochromatin packaging (Fedoroff, 2012; O'Sullivan et al., 2009). Due to the low expression level, young genes that may be deleterious are more beneficial if located in subtelomeric domains, facilitating the process to gene neo-functionalisation (Brown et al., 2010).

Repeatedly, TE enriched regions are important in assisting the formation of adaptive loci through frequent genomic rearrangements and epigenetic regulation (Fedoroff, 2012). Changes in expression through rapid chromatin remodelling or rapid expansion of the gene family through segmental duplications of the environmental response gene can easily take place in these TE enriched regions (Batada and Hurst, 2007; Brown et al., 2010).

1.12. The avenacin gene cluster – a lesson in genome architecture evolution

Metabolic gene clusters are optimized gene arrangements that form and persist through the dynamics of ever scrambling genomes, at least for as long as they confer a selective advantage. The avenacin gene cluster provides a model for investigation of the assembly and functional significance of this form of genomic

organisation. In this thesis, I address this challenge by investigating: 1) the emergence of the avenacin biosynthetic genes within the *Poaceae* and 2) the variations of the avenacin biosynthesis amongst oat species within the subtribe *Aveninae*.

Chapter 2 - Investigation of avenacin production in the tribe *Aveneae*

2.1. Introduction

2.1.1. Avenacins: antifungal plant defence compounds

The avenacins are specialised antimicrobial defense chemicals, that are produced by oats (Goodwin and Kavanagh, 1948; Goodwin and Pollock, 1954). These compounds are produced in the root tips as part of normal growth and development (Osbourn et al., 1994). They have attracted considerable interest because they confer resistance to the take all fungus, *Gaeumannomyces graminis* var. *tritici* which causes major yield losses on wheat (Bateman et al., 2006). Avenacins were first isolated and named by Maizel and Mitchell (Burkhardt et al., 1964; Maizel et al., 1964).

During the period from 1960-1990, most avenacin related research focused on studies of avenacin content, biochemical structure, and biological activity, mainly in the hexaploid cultivated oat species *Avena sativa*. A genus-wide survey was conducted to assess the distribution, composition and antifungal activities of avenacins (Crombie and Crombie, 1986). The authors reported that avenacin A-1 is only found in high content in the roots of *Avena spp.*, although traces of avenacin A-1 were detected in the closely related species *Arrhenatherum elatius*. Screens for presence of avenacin A-1 in different oat varieties and in other grasses were also carried out by several groups (Gibson and Krasnoff, 1999; Mert-Turk et al., 2005; Thomas et al., 2006). These screens involved assessing the roots of young seedlings for the presence of UV- fluorescent material and high performance liquid chromatography (HPLC) analysis of root extracts. Within the genus *Avena*, only the primitive oat species *A. longiglumis* has been found to be avenacin-deficient (Osbourn et al., 1994) so far.

Most plant metabolic gene clusters are species-specific and evolutionarily recent (Field et al., 2011; Field and Osbourn, 2008; Takos et al., 2011). It also has been speculated that the avenacin biosynthetic gene cluster is a recent genomic innovation formed after the divergence of oats from other grasses (Mugford et al.,

2013; Osbourn, 2010b). In contrast, the DIBOA cluster is regarded as an ancient monocot gene cluster that is still intact in *Zea mays* but has been split in rye and wheat (Sue et al., 2011).

2.1.2. Revisiting the oat phylogeny

The avenacin gene cluster has been characterised in the *A. strigosa* accession S75 (Qi et al., 2004). Most other oat species produce avenacins but it is not known whether the genes are also clustered in these species (Crombie and Crombie, 1986). Furthermore, it is not known how many of the cloned and characterised avenacin biosynthetic genes are present/absent in other related species. Systematic analysis of the avenacin content of a collection of carefully selected *Avena* species will provide a starting point for in-depth molecular genetic and bioinformatics-based investigations of the evolutionary boundaries of avenacin production and the birth of avenacin biosynthesis. The formation of the avenacin gene cluster will then be investigated through examination of the absence/presence and expression profile of the *Sad* genes within avenacin producing species. To investigate the evolution of the avenacin pathway, the basal species of the genus *Avena* and their close relatives will also be screened for the absence/presence of avenacins. A critical step in this analysis is therefore to firstly establish the phylogenetic relationships between species in the sub-tribe *Aveninae* through analyses of molecular markers. This will provide a phylogenetic framework for the investigations of the evolution of the avenacin biosynthesis.

Molecular techniques for reconstruction of the oat phylogeny

Avena species belong to the tribe *Aveneae*, one of the major lineages of the *Pooideae* (Figure 2.1) (Loskutov, 2008). The *Aveneae* contain two basic genome types: the A and the C genomes (Figure 2.2). The B and the D genomes present in some tetraploids and hexaploids were found to be derived variants of the A genomes (Badaeva et al., 2010b; Loskutov, 2008). Minor structural variations in the karyotypes have also been found among the diploid oat species. Thus the A and the C genomes of the diploid *Avena* species have been further designated as Cp, Cv and Ac, Ad, Al, Ap, and As, according to their respective karyotype morphologies (Figure 2.2) (Loskutov, 2008). The genomes of most oat species consist of a single type of, or a combination of the A, C, B or D genomes. The only exception is *Avena macrostachya*, an out-crossing species, which has a CmCm tetraploid genome that is closely related to the Cp genomes (Badaeva et al., 2010b). The *Avena* genus is closely related to the genera *Helictotrichon,*

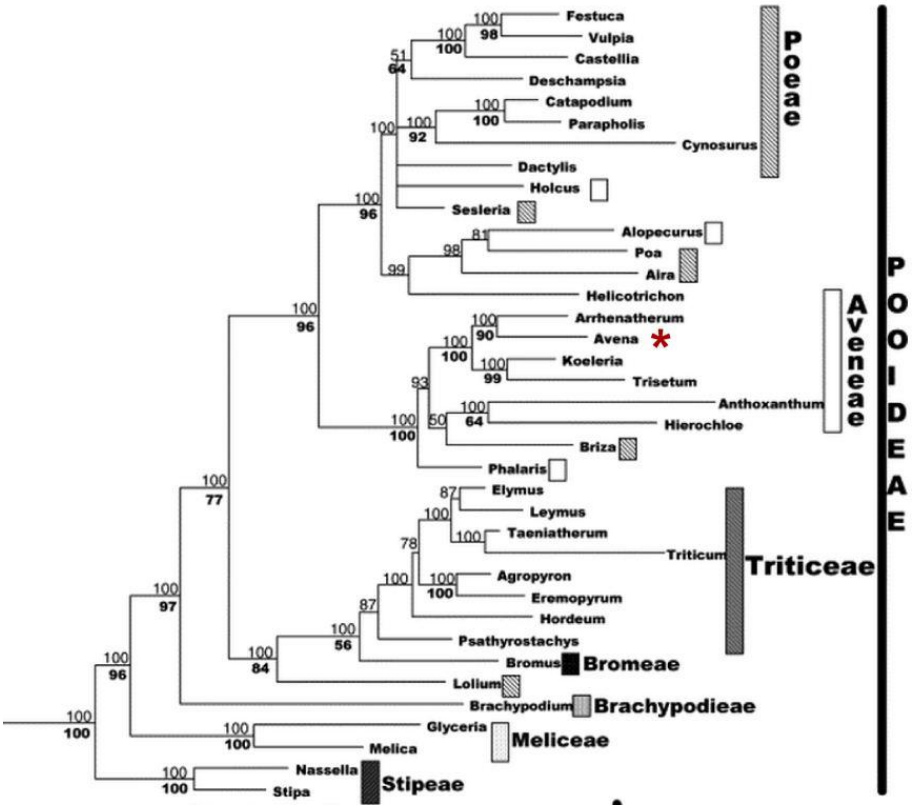


Figure 2.1: Clade of *Pooideae* from the Bayesian consensus estimation using *matK*, *rbcL* and *trnL-F* sequences reproduced from Bouchenak-Khelladi et al. (2008). The *Avena* clade is marked with an asterisk.

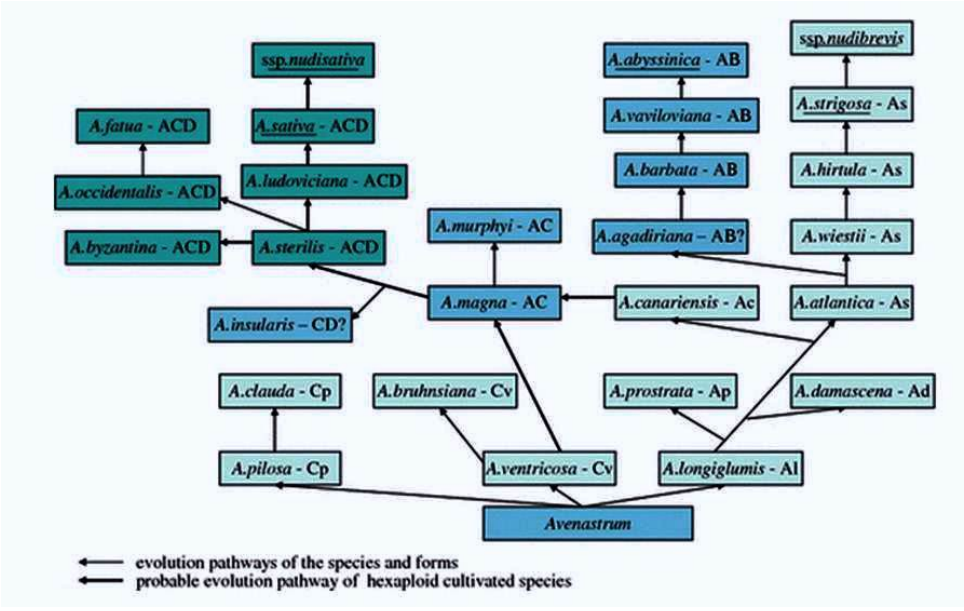


Figure 2.2: Phylogenetic relationships of *Avena* species. The evolution of *Avena spp.* based on the demographic history of oats. Reproduced from Loskutov (2008).

Arrhenatherum (Figure 2.1) and *Sesleria*.

Recent studies of the evolution of the tribe *Aveneae* have focused mainly on cytogenetics, comparative studies of karyotypes and analyse s of molecular markers (Badaeva et al., 2010a,b; Drossou et al., 2004; Fu and Williams, 2008; Morikawa and Nishihara, 2009; Nikoloudakis and Katsiotis, 2008; Nikoloudakis et al., 2008; Peng et al., 2010a,b,c; Rodionov et al., 2005; Winterfeld et al., 2009). However, there are anomalies in the species phylogenies generated in these studies due to: 1) differences in the choice of species, and 2) the different methods of investigation used. For example, ribosomal internal spacer 1 and 2 (ITS1, ITS2) (Peng et al., 2010a; Quintanar et al., 2007), intergenic spacer region (IGS) (Nikoloudakis et al., 2008) and *FLORICAULA/LEAFY* intron2 (FL int2) (Peng et al., 2010c), are nuclear molecular markers, while the *maturase K* (*matK*) gene (Peng et al., 2010b) and *trnL-trnF* intergenic spacers (Peng et al., 2010b; Quintanar et al., 2007; Saarela et al., 2010) are examples of plastid markers that have been used for phylogenetic studies of the tribe *Aveneae*. Furthermore, each marker has its own idiosyncrasies. Because the ITS2 spacers of As genomes are highly similar, the phylogenetic relationships among these species were not well resolved using this marker (Nikoloudakis et al., 2008; Peng et al., 2010a; Quintanar et al., 2007). In the polyploids, the C subgenomes have undergone losses of the 45S rDNA region, so causing the AC tetraploid and ACD hexaploids to cluster with A genome oats in ITS studies (Badaeva et al., 2010a). Phylogenetic studies of plastid markers, on the other hand, only survey the maternal genealogy of the evolution of polyploid oats (Badaeva et al., 2010b; Nikoloudakis and Katsiotis, 2008; Peng et al., 2010b). Therefore, inclusion of both bi-parental markers and maternal markers is important to generate a more unbiased and accurate species phylogeny.

The latest evolutionary model of oats infers that *A. macrostachya* (CmCm tetraploid) was the ancestral *Avena* species, subsequently diverging to give the A and C genome diploids via chromosomal rearrangements (Badaeva et al., 2010a). Subsequent speciation events of the A and C genome diploids were resulted from further chromosomal rearrangements (Badaeva et al., 2010a). Tetraploids originated via hybridisation events between diploid oats, and hexaploids were subsequently formed via hybridisation events between diploid and tetraploid oat species (Loskutov, 2008). The A genome representatives *Avena canariensis, Avena damascena, Avena longiglumis*, and *A. strigosa* all show conserved genomic features, supporting the common origin of the A genome species (Shelukhina et al., 2008a). *A. longiglumis* was considered likely to be the most ancient species of the A genome oats because its karyotypes were most symmetrical compared to those of the other species within the A genome

group (Loskutov, 2008; Shelukhina et al., 2008a). Variation in genome structure captured by C-banding and fluorescent in situ hybridization (FISH) of 5S rDNA probes amongst closely related A genome diploids, C genome diploids and hexaploids supported this evolutionary model (Badaeva et al., 2010a,b; Nikoloudakis and Katsiotis, 2008; Shelukhina et al., 2008a,b). These findings were contradicted, however, by studies with molecular markers (Drossou et al., 2004; Peng et al., 2010a,b,c), which supported the evolutionary order of *Avena L.* as being Ap→Al→Ad→Ac→As. Phylogenetic studies of molecular markers, cytogenetic analysis (Badaeva et al., 2010a), amplified fragment length polymorphism (AFLP) analyses (Fu and Williams, 2008) and consensus chloroplast simple sequence repeat (ccSS) marker diversity analysis (Li et al., 2009) also suggest different species relationships amongst the A genome species. Comparative cytogenetic studies of tetraploids and hexaploids inferred that *A. longiglumis* was the maternal donor of AC tetraploids and that *A. damascena* and *Avena ventricosa* were likely parents of the polyploid *Avena fatua* (Nikoloudakis and Katsiotis, 2008; Peng et al., 2010b). Morphogeographic data (Baum, 1977) suggested that the ancestral oat species originated in the Mediterranean region, and that domestication of accessions such as *Avena sativa* and *Avena barbata*, had led to the spread of these species world-wide. The oat phylogeny is therefore of great interest due to the agricultural importance of *A. sativa* and the complex inter- and intra-ploidy relationships.

The ancestral CmCm genome of *A. macrostachya* possesses a highly symmetrical karyotype. Extensive genome rearrangement of the CmCm genomes of *A. macrostachya* is believed to have given rise to the diploid Cv and Cp genome variants exhibiting marked karyotype asymmetry. In contrast, the A genome divergence from *A. macrostachya* only involved several chromosomal rearrangements which did not disturb the overall genomic composition (Badaeva et al., 2010a). The C genome species also possess "diffuse heterochromatin", formed via the amplification of highly repetitive and C genome specific nucleotide sequences (Shelukhina et al., 2008b). The Cv genome divergence from the C genomes is believed to be due to the loss of the major nucleolus organizer (NOR) on chromosome 3 and significant changes in karyotype morphology via chromosome translocation or inversion (Badaeva et al., 2010a). A further investigation of the phylogenetic relationships of *Avena* and other members of the oat tribe showed *A. macrostachya* to be closely related to *Helictotrichon jahandiezii*, and *Arr. elatius* to be more closely related to the oat species possessing the A genomes than C genomes (Winterfeld et al., 2009).

Here, phylogenetic analysis of a selection of DNA barcodes (CBOL-Plant-Working-Group et al., 2009) (ITS2, *matK*, *trnL-F* spacer) was

carried out to shed further light on the relatedness between *Avena spp.* and other members of the sub-tribe *Aveninae*. Making use of the previously accumulated molecular marker data from various phylogenetic studies (Badaeva et al., 2010b; Nikoloudakis and Katsiotis, 2008; Nikoloudakis et al., 2008; Peng et al., 2010a,b; Quintanar et al., 2007; Saarela et al., 2010; Winterfeld et al., 2009) the evolutionary hierarchy of the *Avena spp.* was revisited.
An analysis of the distribution of the avenacin biosynthesis pathway across the genus *Avena* and other monocot species was carried out to identify candidate species for further investigations of *Sad* gene cluster conservation. In preliminary experiments, roots of seedlings were surveyed by screening for bright-blue fluorescence under UV illumination (indicative of the presence of avenacin A-1). Avenacin content was then further analyzed by thin layer chromatography (TLC) and liquid chromatography/mass spectrometry (LC-MS) analysis of root extracts.

2.2. Materials and Methods

2.2.1. Oat phylogenetic analysis

Sequence retrieval of molecular markers

Sequences of the *matK* and, *trnL-F* spacer, and ITS molecular markers were obtained from previously published or analysed datasets (Nikoloudakis et al., 2008; Peng et al., 2010a,b; Quintanar et al., 2007; Rodionov et al., 2005; Saarela et al., 2010; Winterfeld et al., 2009) (details in Appendix table 2.1). Where possible any partial sequences were replaced with a complete or longer sequence (if these were available in NCBI Genbank (www.ncbi.nlm.nih.gov/genbank/)). For the ITS markers, complete sequences of ITS1, 5.8S rDNA, and ITS2 of the same species were joined together to generate a concatenated ITS1-5.8S-ITS2 sequence for alignment, if complete ITS1-5.8S-ITS2 sequence were not otherwise available. Redundant sequences were removed from the analyses before construction of multiple sequence alignments.

Multiple sequence alignment and phylogenetic tree estimation

Preliminary alignments of ITS1-5.8S-ITS2, matK, and trnL-F spacer necleotide sequences were made separately using MUSCLE 3.6 (Edgar, 2004) and the alignments manually refined using BioEdit (Hall, 1999). Partial sequences were removed from the alignment. Final alignments were created by removal of all

columns containing gaps. Each multiple sequence alignment was input into the FindModel software (Posada and Crandall, 2001) for identification of the best substitution model, according to the Akaike information criterion (AIC) value, to be used for phylogenetic tree construction. Phylogenetic trees were constructed using RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) with the designated model from FindModel (Posada and Crandall, 2001) . In the phylogenetic trees constructed using RAxML 7.0.4 (Stamatakis, 2006) , the best-likelihood tree was obtained by constructing 100 maximum likelihood trees. Then 10,000 bootstrapped trees were constructed and the bootstrap values were mapped to the topology of the relevant best likelihood tree. In the phylogenetic trees constructed using MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001), 50 consensus majority trees were obtained from MCMC analyses from 100,000,000 samples of two parallel runs with the burnin factor of 0.25 and sampled every 1,000 bootstraps until the posterior probability of samples converged to <0.05. The 50 majority consensus tree was then summarised from the 75,000,000 ratined MCMC samples.
Because the ITS1-5.8S-ITS2 alignment contained two regions that are under different selective constraints (ITS1 and ITS2 intergenic non-coding regions will be under relatively neutral selection compared to the RNA coding 5.8S rDNA region), the sequence alignment was partitioned into independent regions. The ITS1 and ITS2 regions were separated from the 5.8S rDNA regions to create two partitions in RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) MCMC analysis.

Supertrees and total evidence (TE) tree construction

The individual phylogenetic trees estimated from the three molecular markers investigated were combined to estimate a more comprehensive phylogeny of oats using CLANN (Creevey and McInerney, 2005), according to the user instructions. Both average consensus and heuristic search approaches were used to identify the best supertrees.
The TE trees were estimated from a concatenated sequence alignment of *matK*, *trnL-F* and ITS markers. Only those species that contained complete sequences for all three molecular markers were included in the TE tree analyses. The supermatrix was built by concatenating the individual alignment of *matK*, *trnL-F* and ITS markers. Because the three molecular markers may be under very different selective pressures, phylogenetic trees were constructed in RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) both with and without partitioning the supermatrix. The supermatrix was separated

into four partitions (*matK*, *trnL-F*, ITS1 and ITS2, and 5.8S rDNA). Trees generated with and without partitioning were compared to one another to evaluate whether these molecular markers were likely to be evolving in very different ways and the impact of partitioning the supermatrix on the tree topologies.

2.2.2. Avenacin screens

Plant material

In each experiment, seeds of each of the accessions listed in Table 2.1 were dehusked and then sterilized by washing first in 5% sodium hypochlorite solution and then in distilled water. The seeds were then placed on moist filter paper in Petri dishes and the dishes sealed with parafilm. The seeds were kept at 4°C for 7 days and then transferred to a growth cabinet for germination (16hr/8hr day/night cycle at 22°C). Seedlings were examined for root fluorescence under

Species	Accession no./lab ref. no	Genome designation
Avena strigosa S75	S75	AsAs
Avena strigosa S75 *sad1* mutant	S75 *sad1* F5 109	AsAs
Avena prostrata	Cc7191 Cs30/1(1)	ApAp
Avena damascena	Cc7258	AdAd
Avena canariensis	Cc7173	AcAc
Avena longiglumis	Cc4719	AlAl
Avena pilosa	JIC2087	CpCp
Avena clauda	180	CpCp
Avena ventricosa	179	CvCv
Avena fatua	A. fatua 2701	AACCDD
Avena sterilis	ISL399(32)	AACCDD
Brachypodium distachyon	B200	
Wheat	Variety Riband and Shamrock	

Table 2.1: Table of species used in the avenacin screen

UV illumination using a transilluminator (320 nm) and photographs taken. *A. strigosa* accession S75 (WT) and *A. strigosa sad1* mutant accession 109 were included as positive and negative controls.

Time (minutes)	% acetonitrile
0	20
3	25
20	50
30	80
32	80
33	20
45	20

Table 2.2: Acetonitrile gradient used for separation of avenacins on a LunaC18(2) column (Phenomenex).

Metabolite analyses of root extracts

For metabolite analysis seeds were germinated as above (with two sets of 10 seeds per replicate). The roots of 5-day old seedlings from each Petri dish were cut off and ground in liquid nitrogen. The powdered root material was extracted with 75% methanol and the extract concentrated to give the equivalent of 1 mg of fresh root/10μl 100% methanol. For TLC analysis, 100 μl of root extract in 75% methanol was dried down by vacuum centrifugation at room temperature and the pellet then resuspended in 20 μl of 100% methanol. The root extracts were then loaded on to a silica gel 60 (MERCK®) thin layer chromatography (TLC) plate and the TLC was developed in chloroform:methanol:water at a ratio of 13:6:1. The presence/absence of the major and minor fluorescent avenacins, A-1 and B-1, was visualised under UV illumination. Supernatants were transferred to glass vials for LC-MS analysis, which was carried out by the JIC Metabolite Services. Samples were analysed on a Surveyor HPLC system attached to a DecaXPplus ion trap mass spectrometer (Thermo®). The avenacins were separated on a 100 x 2 mm 3μ Luna C18(2) column (Phenomenex) using a gradient of acetonitrile versus 0.1% formic acid in water (Table 2.2), run at 30°C and 300 μL.min-1. All four forms of avenacin were detected by UV absorbance and mass spectrometry. UV spectra were collected from 190-600nm, while positive mode electrospray MS data were collected from m/z 150-2000. The four forms of avenacin surveyed in the LC-MS analyses were quantified using the LC-MS analysis software package Finnigan™ Xcalibur®. The relevant avenacin was detected by screening the full mass spectrum with the mass/charge ratio of either the hydrogen adducts or the sodium adducts (Table 2.3). Then the mass fragmentation pattern (MS2) was examined for validation of identity. The relative intensities of the different forms of avenacin were calculated from the area of the peaks in the liquid chromatography spectrum.

Avenacin adducts	A-1	A-2	B-1	B-2
+H	1094.5	1065.5	1078.5	1049.5
+H -1 sugar	943.5	903.5	916.5	887.5
+H -2 sugars	770.5	741.5	754.5	725.5
+H -3 sugars	638.4	609.5	622.4	593.5
+Na	1116.5	1087.5	1100.5	1071.5
+Na -1 sugar	954.5	925.5	938.5	909.5
+Na -2 sugars	792.5	763.5	776.5	747.5
+Na -3 sugars	660.5	631.5	644.5	625.5

Table 2.3: Mass/charge ratio of avenacins and avenacin adducts

2.3. Results of the oat phylogenetic analysis

The phylogeny of the *Aveninae* sub-tribe was constructed using the combined phylogenies obtained from ITS, *trnL-F* and *matK* data according to previous recommendations on molecular markers (CBOL-Plant-Working-Group et al., 2009; Hollingsworth, 2011; Peng et al., 2010b). *matK* has been rated as one of the most effective markers of discriminative power while the *trnL-F* and ITS sequences provided high universality and species coverage. ITS sequences were included to infer both paternal and maternal origins, whereas the *matK* and *trnL-F* markers reflect maternal origins only. It has been shown previously that inclusion of ITS sequences in phylogenies built from both the *matK* and *rbcL* barcodes increases the discriminative power of species by 20% (Hollingsworth, 2011).

2.3.1. Maternal molecular markers – *matK* and *trnL-F* spacer.

trnL-F has been reported as a stable chloroplast marker (Peng et al., 2010b) while *matK* is a rapidly evolving sequence capable of discriminating between recently diverged species. *matK* sequences were comparatively scarce among our target species compared to *trnL-F*. In contrast, *trnL-F* sequences were more widely available in other *Aveninae* species.

The *trnL-F* results estimated in RAxML 7.0.4 and MrBayes 3.2.1 contained highly consistent groupings (Table 2.4, Figure 2.3a and b), with both showing that the A genome and C genome oats were separated into two clades. *Arrhenatherum elatius* was found to be closely related to the A genome oats (bootstrap value = 60), whereas *A. macrostachya* and four *Helictotrichon spp.* (*Helictotrochon convolutum, Helictotrichon sempervirens, Helictotrichon filifolium* and

Summary of *matK*, *trnL-F* and ITS phylogenetic estimation					
Phylogenetic software	No. of sequences	Alignment length (bp)	lnL	(Mean) α	Substitution model
trnL-F					
RAxML	134	436	-1280.31	0.51	GTR + Γ
MrBayes	134	436	-1553.75	0.07	GTR + Γ
matK					
RAxML	72	1393	-3770.48	0.49	GTR + Γ
MrBayes	72	1393	-3954.83	0.38	GTR + Γ
ITS1-5.8S-ITS2					
RAxML	331	445	-6630.04	(ITS)1.14, (5.8S rDNA)5.27	GTR + Γ 2 partitions
MrBayes	331	445	-7243.61	(ITS1)1.55, (5.8S rDNA)0.15, (ITS2)1.30	GTR + Γ 3 partitions

Table 2.4: Summary of phylogenetic trees of *matK*, *trnL-F*, and ITS1-5.8S-ITS2 markers constructed in RAxML 7.0.4 and MrBayes 3.2.1

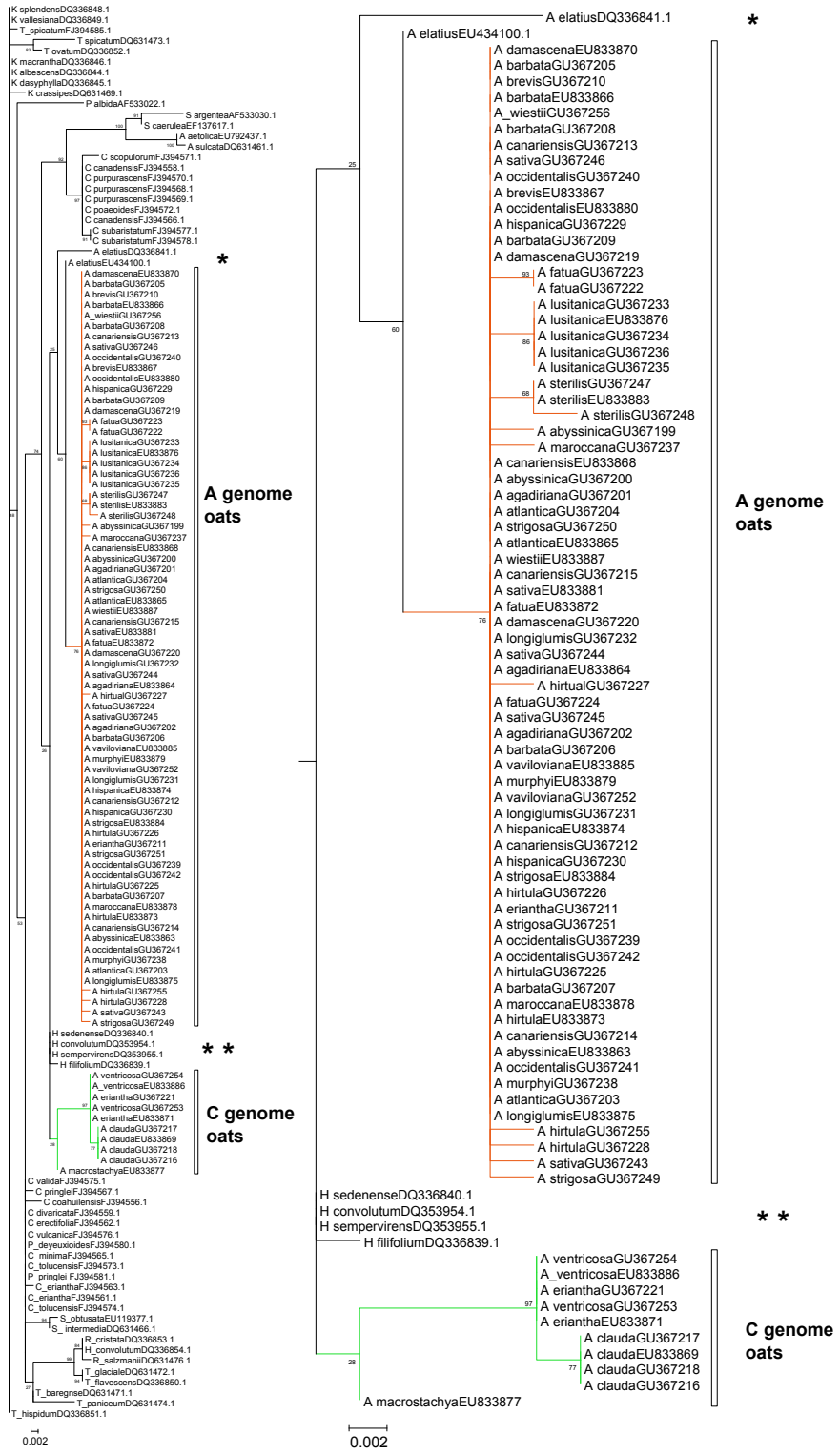


Figure 2.3: *trnL-F* tree of the sub-tribe *Aveninae* estimated in a) RAxML 7.0.4. The *Avena* subclade is displayed on the right. Branches leading to A and C genome oats are highlighted in orange and green respectively. *Arr. elatius* (*) and *Helictotrichon spp.* (* *) are indicated with asterisks. Bootstrap support for branches (from 10,000 replicates) are indicated. The original tree file is in the Appendix (Tree 2.1). The Bayesian estimation of *trnL-F* is in the Appendix (Tree 2.2).

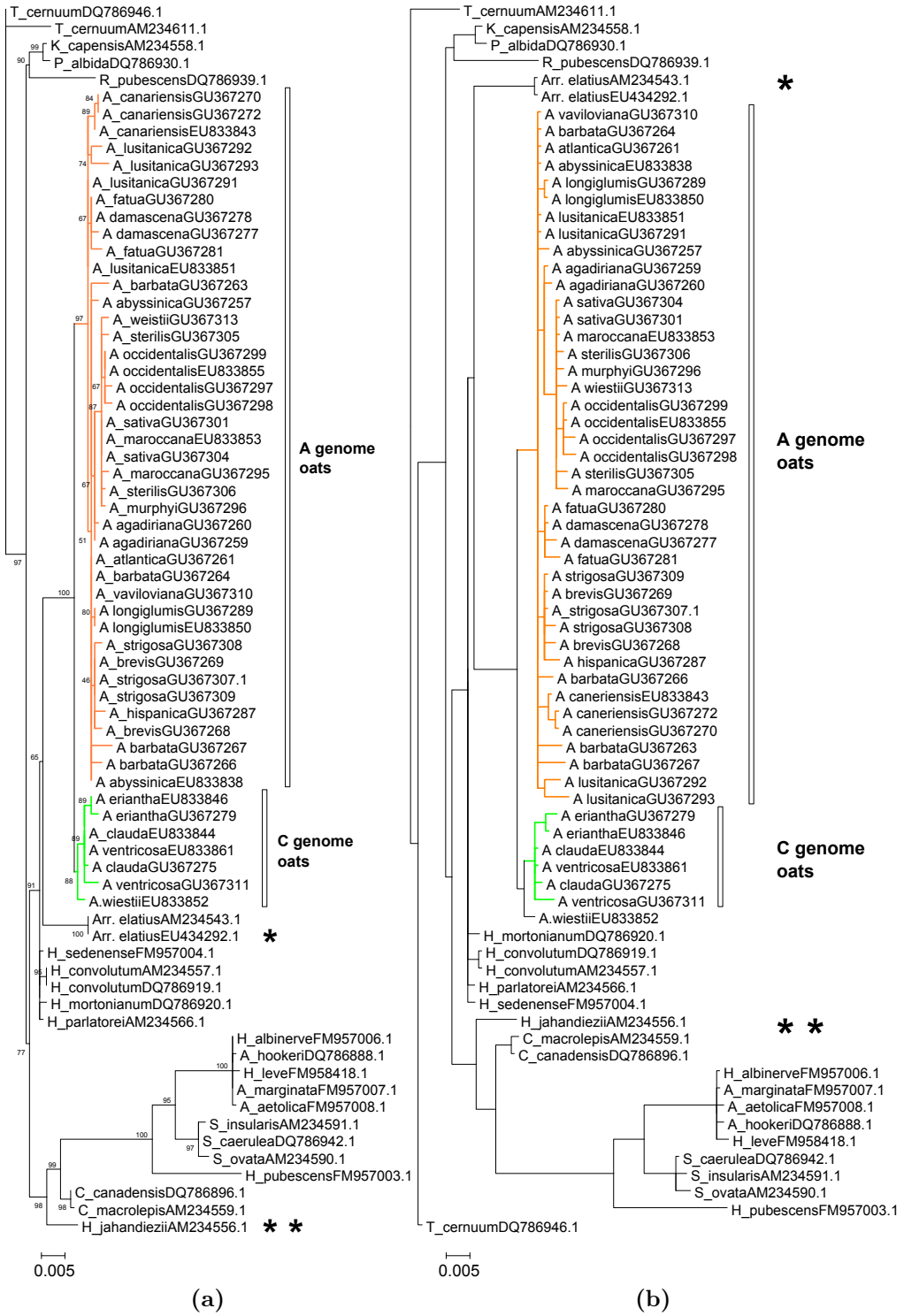


Figure 2.4: *matK* trees of the sub-tribe *Aveninae* estimated in a) RAxML 7.04 and b) MrBayes 3.2.1. Branches leading to A and C genome oats were highlighted in orange and green respectively. *Arr. elatius* (*) and *H. jahandiezii* (* *) are indicated with asterisks. Bootstrap support for branches (from 10,000 replicates in RAxML) are indicated and the posterior probabilities for branches (from 100,000,000 MCMC samples in MrBayes) are not shown. The original tree files are enclosed in the Appendix (Tree 2.3 and 2.4).

Helictotrichon sedenense) were closely related to the C genome oats, consistent with the previous finding (Winterfeld et al., 2009). The sequence representatives of the modified A genomes (*A. longiglumis*, *A. canariensis*, and *A. damascena*) were not divergent from those of As genome species such as *A. strigosa*, *Avena brevis* and *Avena hirtula*, reflecting that the *trnL-F* sequences did not contain enough inter-specific variations to resolve the oat phylogeny within the A genome species. Interestingly, the polyploids all fell into the As genome clade, even though the ACD hexaploids also contain the C sub-genomes, suggesting that their maternal donor possessed an A genome. In the C genome clade, *Avena eriantha (pilosa)* had previously been reported to be closely related to *A. clauda* (both species being assigned as CpCp diploids). Here it grouped instead with *A. ventricosa*, which possesses the Cv genome (bootstrap value = 97).

The phylogenetic estimatess in RAxML 7.0.4 and MrBayes 3.2.1 exhibited consistent groupings and the overall topologies are highly similar. The *matK* results (Table 2.4, Figure 2.4a and b) also showed that the A genome and C genome oats were separated into two clades. Distinct *matK* sequences from *A. wiestii* grouped with both the A and the C genome oats, resulting in conflicts in assigning the location of this species within the *matK* tree. The *matK* tree also indicated that *Arr. elatius* was more closely (bootstrap value = 65) related to the *Avena spp.* than to the genus *Helictotrichon*, an outcome that differs from the previous ITS analysis of Quintanar et al. (2007) and our *trnL-F* analysis. *A. ventricosa* clustered with *A. clauda* within the C genome oat clade, more distantly with *A. eriantha* (bootstrap value = 89) in the *matK* tree (Figure 2.4a and b).

Nonetheless, the *matK* phylogeny exhibited better resolution amongst the A genome oats compared to the *trnL-F* trees. The A genome clade indicated that the As genome species *A. hispanica, A. brevis*, and *A. strigosa* were clustered together. *A. damascena* was clustered with *A. fatua*, suggesting its contribution as the maternal donor of the hexaploid. *A. agadiriana, A. sativa, A. maroccana, A. murphyi, A. occidentalis* and *A. sterilis* were also clustered together, suggesting a common origin.

2.3.2. Nuclear molecular markers – ITS sequences

ITS sequences were the most widely available molecular marker surveyed. Because multiple ITS sequences have been reported in all oat species, presumably due to sequence heterogeneity within the rDNA tandem arrays, all ITS entries for each oat species were included in the multiple sequence alignment in an attempt to capture the intra-genomic ITS variations in the species tree.

The corresponding regions of ITS1, 5.8S rDNA and ITS2 were identified in the refined alignment. In the phylogenetic tree construction under the GTR + Γ substitution model, the sequence alignments of the ITS regions and 5.8S rDNA regions were treated separately, with individual parameters for base frequency, substitution rate heterogeneity, and transition/transversion ratio, etc.

In both ITS trees, the A genome oats clustered within a broader clade of C genome oats (bootstrap value = 78). *A. macrostachya* was located at the base of the *Avena spp.* clade, consistent with the previous analysis (Badaeva et al., 2010a). Interestingly, ITS sequences of *A. weistii* and *A. atlantica* were found to be clustered with *Calamagrostis spp.* and *Helictotrichon spp.* rather than with the A genome oats, an observation that has not been reported in previous studies (Quintanar et al., 2007; Saarela et al., 2010).

In the ITS tree estimated using RAxML 7.0.4 (Figure 2.5a), *Sesleria spp.* (bootstrap value = 38) were more closely related to *Avena spp.* than to *Arr. elatius* (bootstrap value = 35) and *H. jahandiezii* (bootstrap value = 62), consistent with previous findings with low bootstrap support (Quintanar et al., 2007).

In the A genome oat clade, most species were located in unresolved polytomies (Figure 2.5b), suggesting they all possess highly similar ITS sequences. Regardless of the elimination of the C subgenome 45S rDNA regions observed in hexaploid oats (Badaeva et al., 2010a), ITS sequences of *A. fatua* mainly clustered with those of *A. ventricosa* (bootstrap value = 51) and some ITS sequences of *A. sterilis* and *A. maroccana* were grouped closer to the C genome oats than to the A genome oats (Tree 2.5 and 2.6 in the Appendix). Consistent with all previous ITS analyses (Peng et al., 2010a; Quintanar et al., 2007; Saarela et al., 2010), the phylogenetic relationship of the A genome oats was poorly resolved in the ITS tree, as for *trnL-F*. The A genome variants *A. longiglumis*, *A. canariensis*, *A. damascena* and *A. prostrata* ITS sequences were grouped with those of other A genome species. The ITS sequences of *A. prostrata* were more closely related to the ITS sequences of *A. damascena* and formed a clade with relatively long branches (bootstrap value =100) in both ITS trees (Figure 2.5), while the ITS sequences of *A. canariensis* and *A. longiglumis* were indistinguishable from those of other A genome oats. Thus the phylogeny of the A genome oats could not be readily inferred. Nonetheless, the ITS tree did not provide any evidence to support the proposed evolutionary model of Ap→Al→Ad→Ac→As (Drossou et al., 2004) because the phylogenetic estimations using molecular markers did not infer the genome type of internal nodes.

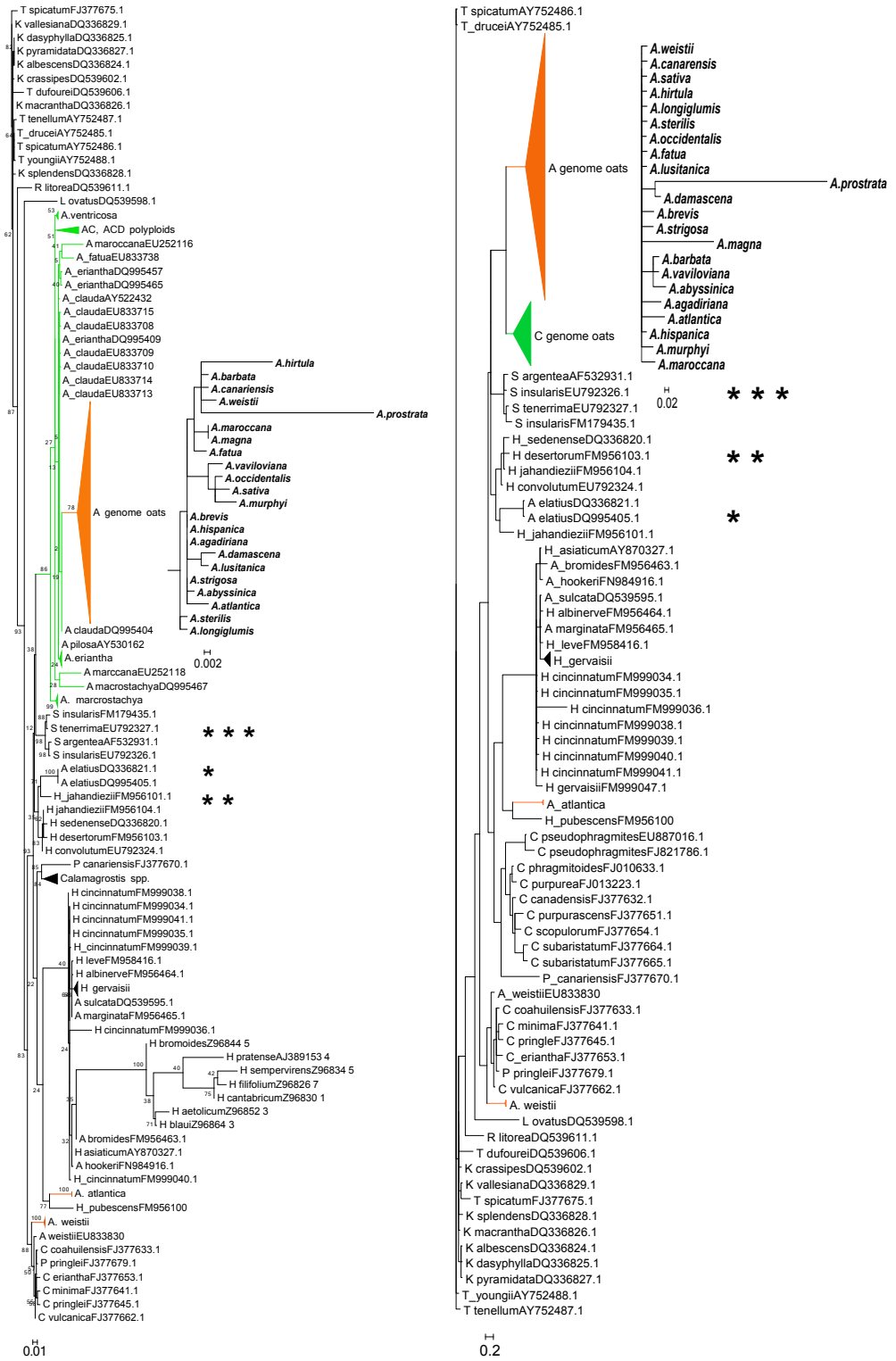


Figure 2.5: ITS trees of the sub-tribe *Aveninae*. Phylogenetic trees of ITS1-5.8s-ITS2 estimated in a) RAxML 7.0.4 and b) MrBayes 3.2.1. The summarised topology of the A genome oats is shown on the adjacent subtree. Branches leading to A and C genome oats are highlighted in orange and green respectively. *Arr. elatius* (*), *Helictotrichon spp.* (* *) and *Sesleria spp.* (* * *) are indicated with asterisks. Bootstrap support for branches (from 10,000 replicates in RAxML) are indicated and the posterior probabilities for branches (from 75,000,000 samples in MrBayes) are not shown. The original trees are enclosed in the Appendix (Tree 2.5 and 2.6).

2.3.3. Similarities and differences between the phylogenies generated using different markers

The phylogenetic trees constructed using the ITS, *matK* and *trnL-F* spacer sequences all showed that the C genome and A genome oats are separated (with the exception of Figure 2.5). *A. macrostachya* is located at the base of the C genome oats in the *trnL-F* and ITS trees. In the *matK* and *trnL-F* trees, all polyploids were grouped with A genome diploid species, which are likely to be the respective maternal donors. The ITS sequences of the AC and ACD polyploid species clustered with both A and C genome oats, suggesting that the C genome oats are the likely parental donors of polyploids. However, no particular C genome species were implicated by the phylogenetic trees for the three molecular markers as the ancestral donors of the polyploid species. The *matK* and *trnL-F* trees suggest that *Arr. elatius* is the closest relatives of the *Avena spp.*, but *Sesleria spp.* were inferred to be the closest relatives in the ITS analyses. In general, the molecular marker trees possess low bootstrap supports (<80) to infer significant interspecific relationships within *Avena spp.*, especially amongst A genome oats.

2.3.4. Supertree of the sub-tribe *Aveninae*

The three sets of phylogenetic analyses (*matK*, *trnL-F* spacer and *ITS*) revealed different phylogenies for the *Aveninae*. The datasets for the *matK*, *trnL-F* spacer and *ITS* markers consisted of sequences from overlapping but not identical groups of species, thus limiting the scope of phylogenetic estimations using consensus tree or total evidence methods that require the use of all three molecular markers for each species. In an attempt to obtain a comprehensive oat phylogeny, supertree methods were employed to summarized the source trees (reviewed in Bininda-Emonds (2004)).

The phylogenetic trees estimated from *matK*, *trnL-F* spacer and *ITS* sequences were input into CLANN (Creevey and McInerney, 2005) for supertree construction. Both average consensus and heuristic search approaches were used to obtain the best supertree estimate. Supertrees were constructed first using source trees of *matK* and *trnL-F* spacer, followed by all three source trees. Phylogenetic trees estimated using RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) were separated during supertree construction to reduce tree topology incongruities related to the differing tree building algorithms. *Trisetum spicatum* was used as an outgroup for rooting in all cases.

All the resulting CLANN supertrees (Creevey and McInerney, 2005) contained very poor species groupings (data not shown) and bootstraps analyses suggested that all supertrees had very low bootstrap support. In general, *Avena spp.* were scattered throughout the trees and no known phylogenetic groupings could be observed within them.

In an attempt to improve supertree construction, one single tip per species was retained in the source trees (Tree 2.7-2.8 in the Appendix) while the redundant tips of the same species were removed from the source trees using the Phylip retree program (Felsenstein, 1989). For *A. fatua*, *Avena sterilis* and *Avena maroccana*, where the majority of ITS sequences clustered with the C genome diploid oats, only one tip clustering with the C genome oats was kept in the source tree for each of these species. Because polyploids are a well known source of phylogenetic conflicts, another set of supertrees was built from source trees without inclusion of polyploid species in order to enhance supertree inference (Tree 2.9-2.10 in the Appendix).

The two sets of simpler source trees (summarized in Table 2.5) were then input again into CLANN (Creevey and McInerney, 2005) for the construction of average consensus supertrees. The average consensus tree generated by CLANN (Creevey and McInerney, 2005) using the modified source trees still exhibited poor lineage relationships (Tree 2.7 and 2.8 in the appendix). However, heuristic searches carried out on the source trees generated from RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) yielded improved results that are closer to our current understanding of the oat phylogeny from the individual marker studies and from previously published results (Tree 2.9 and 2.10 in the appendix).

2.3.5. Construction of total evidence (TE) trees from core species

In addition to building supertrees, total evidence (TE) trees were constructed from a supermatrix of *matK*, *trnL-F* spacer and *ITS* sequences of species that contained full-length sequences of all three markers. Concatenated sequences for each species were constructed from one *matK*, *trnL-F* spacer and *ITS* sequence obtained from the multiple sequence alignments previously built for phylogenetic tree construction of individual molecular markers (Subsection 2.2.1). Construction of the phylogenetic trees in RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001), both with and without partitioning of the supermatrix, was performed and the resulting topologies were compared. For the former, the supermatrix was divided into a total of four partitions (*matK*, *trnL-F*,

Summary of Supertree source trees			
removal of duplicates			
Tree	ITS	matK	trnL-F
No. of tips	78	42	68
removal of duplicates and polyploids			
Tree	ITS	matK	trnL-F
No. of tips	65	31	58

Table 2.5: Summary of the modified source trees (removal of duplicate sequences and removal of polyploid sequences) input into CLANN

ITS1 and ITS2, and 5.8S rDNA). Only 27 species possessed full-length sequences of all three molecular markers and thus were all included in the supermatrix of length 2274 bp. *Calamagrotic canadensis* was used as the outgroup sequence for rooting the trees.

The tree topologies estimated using RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) were consistent with one another with respect to their topologies and grouping of species (Figure 2.6a, b and 2.7a, b). Partitioning of the supermatrix in the phylogenetic estimations have greatly improved the maximum likelihoods of the TE trees (Figure 2.6 and 2.7). Furthermore, the tree topologies derived from the concatenated sequences were consistent with previously published phylogenies (Badaeva et al., 2010a,b; Drossou et al., 2004; Fu and Williams, 2008; Nikoloudakis and Katsiotis, 2008; Nikoloudakis et al., 2008; Peng et al., 2010a,b,c; Rodionov et al., 2005; Winterfeld et al., 2009). *A. macrostachya* grouped in all cases with C genome diploid oats, suggesting it shared more similarities with the C genomes across the dataset analysed. *A. clauda* and *A. eriantha* were more closely related to one another than to *A. ventricosa* in the trees obtained from non-partitioning (Figure 2.6b and 2.7b) of the supermatrix, while the three C genome species formed a polytomic clade in the tree obtained from partitioning of the supermatrix (Figure 2.6a and 2.7a). *A. wiestii* and *A. atlantica*, which both possess the As genomes, were found to group with other A genome oats but formed a neighboring clade. The species resolution within the A genome oats was low and it was not possible to infer accurate phylogeny amongst the A genome oats due to low bootstrap supports. However, due to the absence of *A. prostrata* in the concatenated trees, the evolutionary model (Figure 2.2) of the A genome oats could not be fully elucidated. In the RAxML TE trees (Figure 2.6a and b), *A. longiglumis* clustered with the AC tetraploids and ACD hexaploids. This grouping agreed with the previously published literature (Nikoloudakis and Katsiotis, 2008; Peng et al., 2010b). However, in the MrBayes TE trees (Figure 2.7a and b), *A. longiglumis* clustered with the As diploids and AB tetraploids,

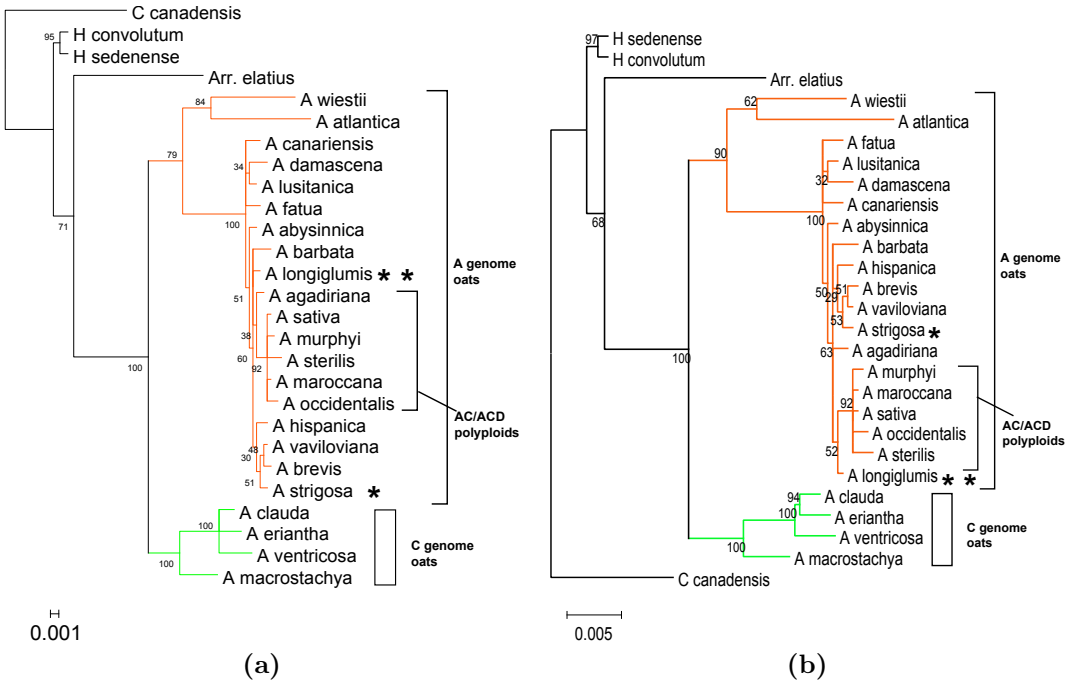


Figure 2.6: TE tree generated using RAxML. phylogenetic tree generated from a supermatrix of *matK*, *trnL-F* spacer and *ITS* sequences in RAxML (a) with partitioning (lnL = -4817.04, α = 0.19, 0.10,0.24, 0.38) and (b) without partitioning (lnL = -4963.16, α = 0.02). Branches leading to A genome and C genome oats are highlighted in orange and green respectively. The lab reference species *A. strigosa* (*) and *A. longiglumis* (* *) is marked with asterisks. Bootstrap support for branches (from 10,000 replicates) are indicated. $2\Delta L = 292$ (degree of freedom = 49. $p<0.01$).The original tree files are in the Appendix (Tree 2.11 and 2.12).

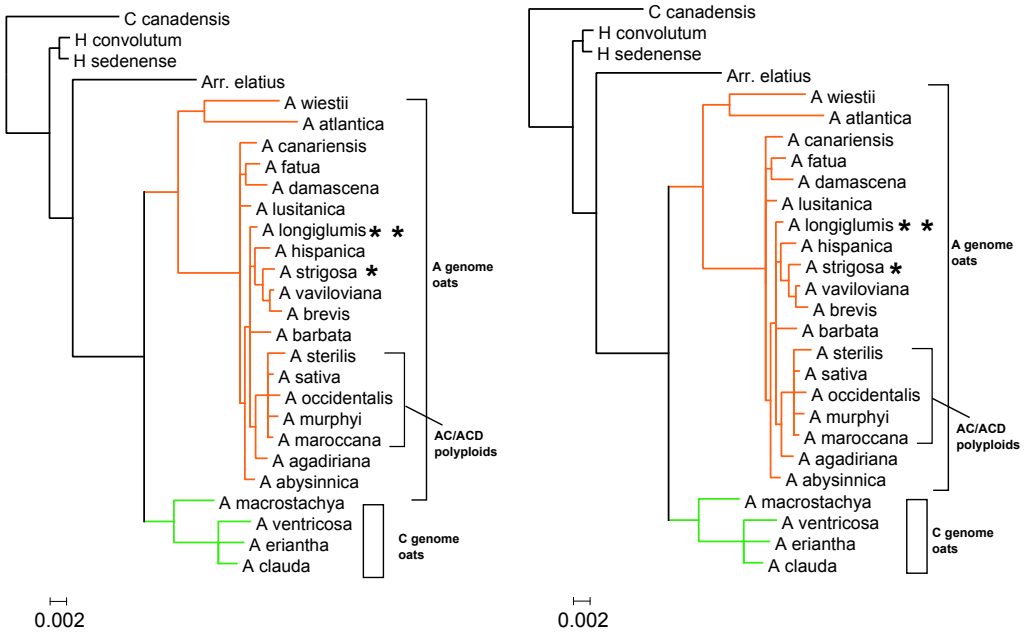


Figure 2.7: TE tree generated using MrBayes. Phylogenetic trees generated from a supermatrix of *matK*, *trnL-F* spacer and *ITS* sequences in MrBayes (a) with partitioning (lnL = -4931.33, α = 78.91, 92.95, 0.81, 99.97) and (b) without partitioning (lnL = -5023.06, α = 0.05). Branches leading to A genome and C genome oats were highlighted in orange and green respectively. The lab reference species *A. strigosa* (*) and *A. longiglumis* (* *) is marked with asterisks. The posterior probabilities for branches (from 100,000,000 replicates) are not shown. $2\Delta L = 184$ (degree of freedom = 49. $p<0.01$) The original tree file are in the Appendix (Tree 2.13 and 2.14).

and exhibited a more distant relationship to the AC tetraploids and ACD hexaploids, which formed an individual clade.
The TE trees provided a foundation for studying the evolution of the avenacin biosynthesis among oat species. Because the published oat phylogenies are in agreement with the oat phylogenetic relationship displayed by the TE trees, the TE tree estimated using RAxML 7.0.4 (Stamatakis, 2006) (Figure 2.6a) was used as the reference oat phylogeny for later investigation of the avenacin biosynthesis evolution.

2.4. Results of the avenacin production screen

Preliminary screens of root fluorescence of seedlings (Table 2.3 in the Appendix) indicated that the bright-blue fluorescence characteristic of avenacin A-1 was present in all of the *Avena spp.* examined, except *A. longiglumis*. However, such fluorescence was absent in other species of the *Aveninae* sub-tribe including *Helictotrichon sedenense*, *Helictotrichon bromide*, *Sesleria caerulea*, *Koeleria vallesiana*, *Koleria crassipes*, *Koeleria pyramidata* and *Briza minor*, consistent with the published root fluorescence screens (Crombie and Crombie, 1986; Goodwin and Kavanagh, 1948). Taking the re-estimated oat phylogeny, the preliminary root fluorescence screen and seed availability into account, metabolite analysis of crude root extracts of ten selected oat species focussing on avenacin production was carried out here to establish the core set of species for further investigations in chapter 3.
Bright blue fluorescence was observed in roots of all oat accessions surveyed except *A. longiglumis* (Figure 2.8). The fluorescence intensity varied among oat species, suggesting that avenacin A-1 content may be species-dependent. The *sad1* mutant *A. strigosa* 109 included in this analysis has a point mutation within the β-amyrin synthase gene and fails to synthesise avenacins (Haralampidis et al., 2001; Papadopoulou et al., 1999). The roots of this mutant were clearly reduced in fluorescence compared to the wild type *A. strigosa* S75 wild type line. The residual fluorescence seen in this mutants is likely to be due to the presence of other fluorescent root compounds such as scopoletin.
The roots of seedlings of the A genome diploid species *A. canariensis, A. damascena* and *A. prostrata* were all strongly fluorescent, suggestive of the presence of avenacin A-1 at levels comparable to those of *A. strigosa*. The hexaploids *A. fatua* and *A. sterilis* (AACCDD genome) were also strongly fluorescent. The C genome diploids *A. clauda, A. pilosa*, and *A. ventricosa* also had some root fluorescence but the fluorescence of *A. pilosa* and *A. ventricosa* was not as intense as that of the A genome species. The roots of seedlings of *A.*

longiglumis had only weak fluorescence, while *B. distachyon* and wheat were non-fluorescent under the conditions used.

2.4.1. Metabolite analyses of root extracts

Root extracts prepared from 5-day old seedlings of the selected oat accessions were analysed using thin layer chromatography (Figure 2.9). The wild-type *A. strigosa* S75 and the *sad1* mutant *A. strigosa* 109 had a clear difference in the intensity of bright-blue fluorescence, indicating that avenacin A-1 was readily detected in the roots of the wild-type but not in the mutant. TLC analysis also clearly showed that diploid A genome oats, except the avenacin-deficient species *A. longiglumis*, produced more avenacin than the C genome species. High level of avenacins A-1 and B-1 were readily detected in the root extracts of the hexaploids *A. fatua* and *A. sterilis*. There was a readily detectable fluorescent metabolite with the same mobility as the avenacin biosynthetic pathway intermediate *N*-methyl anthraniloyl-O-Glc, in both *A. strigosa* S75 and 109. This was also visible in root extracts of *A. pilosa*, *A. damascena* and *A. fatua* at lower levels but not in any other oat species (Figure 2.9). *A. longiglumis* produced no detectable avenacin A-1 or B-1 in the TLC and also a reduced amount of other fluorescent chemicals. No fluorescence was detected in wheat or *B. distachyon* root extracts in the TLC analysis performed here.

Root extracts were then subjected to LC-MS for the detection and semi-quantification of avenacins. Targeted LC-MS sprectra for avenacins and avenacin adducts were detected (Figure 2.10) and their relative quantities were measured. Avenacins A-1, B-1 and A-2 were detected with LC retention times of approximately 18, 16 and 13 minutes in liquid chromatography (± 1 minute), whereas avenacin B-2 was barely detectable in root extracts of most oat species (Appendix Table 2.2). The analysis thus focused on the readily detectable avenacins A-1, B-1 and A-2 for comparison of the avenacin content of the species included in the analysis.

LC-MS analysis revealed that avenacins A-1, A-2 and B-1 are found at relatively high levels in wild-type *A. strigosa* S75 roots but not in the *sad1* mutant *A. strigosa* 109. Semi-quantification of avenacins A-1, A-2 and B-1 showed that avenacin levels are highest in root extracts from *A. strigosa* S75, *A. fatua* and *A. sterilis* at higher levels compared to diploid oat species, with lower level in the C genome diploid species. The ratios of avenacins A-1, A-2 and B-1 were consistent across all oat species analysed; avenacin A-1 was the most abundant of the three (80-90% of total avenacin content), followed by B1 (approximately 10%) and A-2 (approximately 2%).

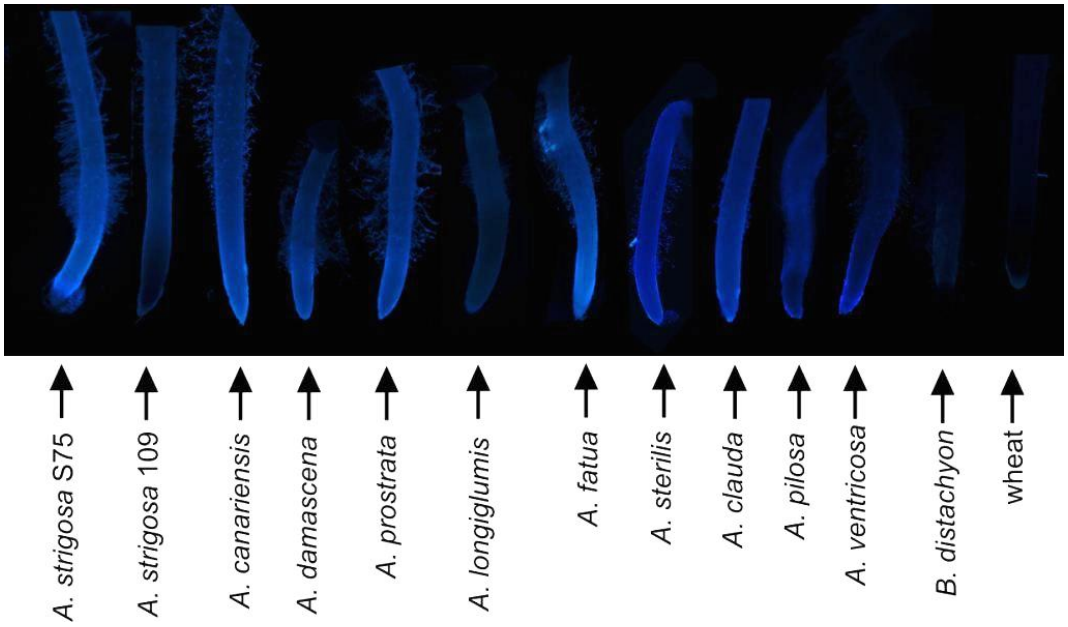


Figure 2.8: Root panel for *Avena spp.* Seedlings with root length of 2-3cm were recorded for avenacin production under UV exposure.

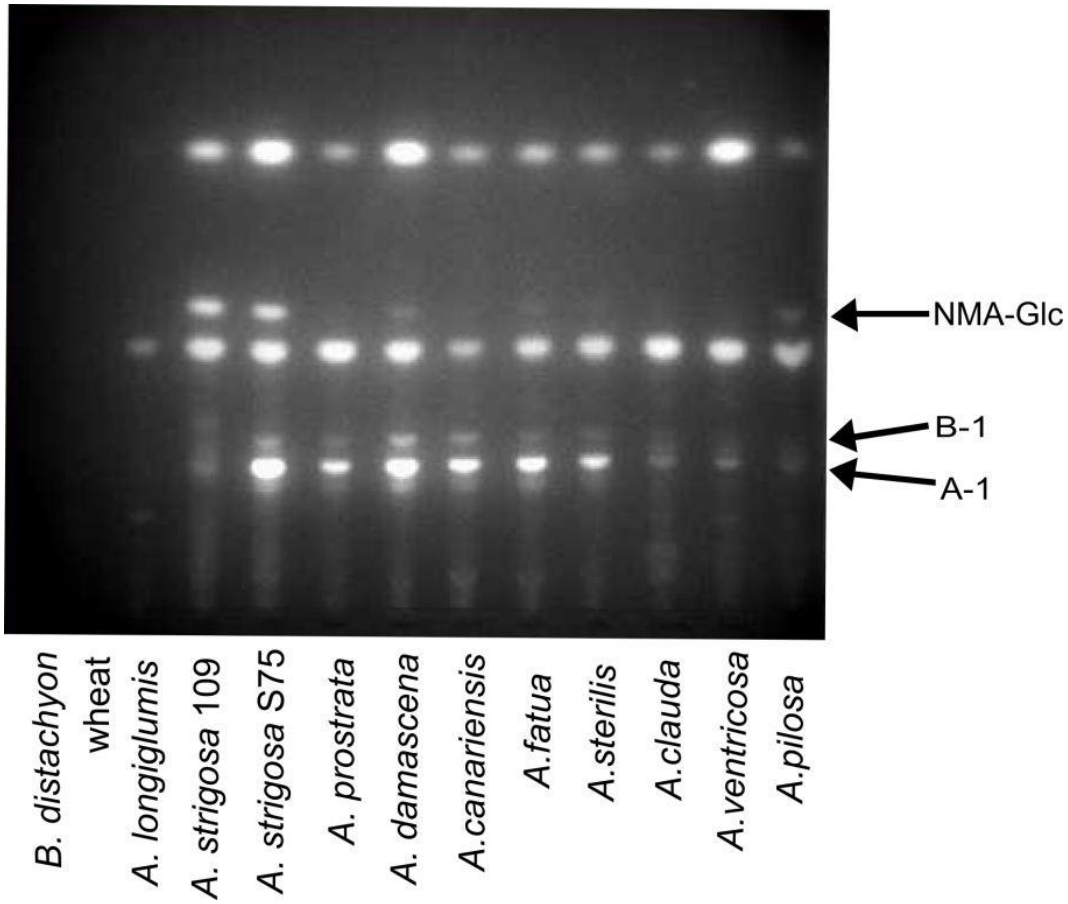


Figure 2.9: TLC analysis of root extracts of *Avena spp.*, *B. distachyon* and wheat. *N*-methyl anthraniloyl-O-Glc, avenacin A-1 and B-1 are arrowed.

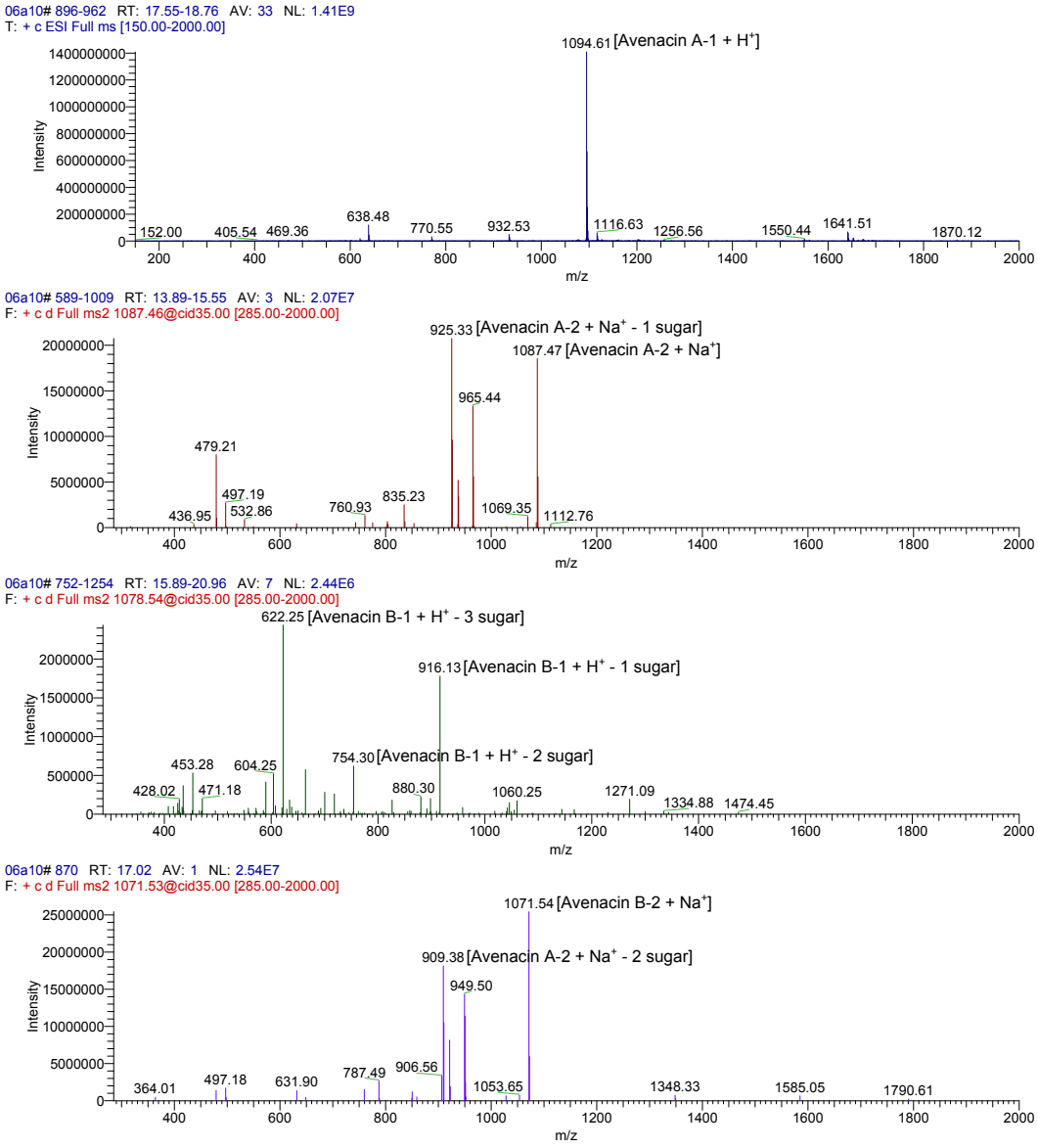


Figure 2.10: Representative LC-MS spectra of avenacin A-1, A-2, B-1 and B-2. The annotated mass sprectra of avenacin A-1 (top panel), A-2 (second panel), B-1 (third panel) and B-2 (bottom panel) detected in *A. strigosa* S75 root extracts. The mass of the relevant avenacins and avenacin adducts (listed in 2.3) are labelled.

2.5. Discussion

2.5.1. Oat phylogenetics

In this chapter the sequences of the molecular barcode markers *matK*, the *trnL-F* spacer and ITS were used to revisit the phylogeny of the *Aveninae* sub-tribe, with a particular focus on the *Aveneae*. *matK*, compared to the other two barcode markers, provided the tree with the highest between-species resolution. The species phylogeny estimated using the *matK* marker was further reinforced by consistentcy with the tree topologies estimated by the *trnL-F* spacer. The phylogenetic trees estimated from the ITS markers had low between-species resolution and poor bootstrap support (Figure 2.5). Most species possess variation within their ITS sequences due to bi-parental inheritance as well as heterogeneity within the ribosomal tandem array yet to be resolved through concrete evolutionary processes, further increasing the complexity of the dataset and posing additional challenges to uncovering their true species phylogeny.

Here, Supertree estimation was performed in order to summarize the phylogenetic information provided by the three set analyses using individual molecular markers. However, supertree approaches were not particularly informative in terms of resolving the inter-specific relationships of the A genome oats. Improved analyses of lineage relationships by ITS sequences may be achieved by using alternative phylogenetic estimation methods such as phylogenetic networks (Peng et al., 2010a) which reflect more complex evolutionary events such as hybridization.

The total evidence approach (TE tree) provided a more plausible reconstruction of the *Avena* species tree. In the TE tree, the A and C genome oat species grouped distinctly and most species were located in groupings consistent with previous publications. However, the full-length sequences of all three molecular markers selected for this phylogenetic study were only fully available for a handful of species in the *Aveninae*, and so the TE tree was restricted primarily to the members of the *Avena* clade. Of note, *A. prostrata*, which has been proposed to be an ancient accession (Drossou et al., 2004) had to be omitted from the TE tree estimation due to a lack of *matK* and *trnL-F* sequences. In the future, experimental determination of the *matK* and *trnL-F* sequences of *A. prostrata* will enable a more comprehensive oat phylogeny to be inferred.

Notwithstanding the fact that the TE tree could not be used to validate the proposed evolutionary order of oat (Ap→Al→Ad→Ac→As), it provided a robust evolutionary framework for further studies of the evolution of the

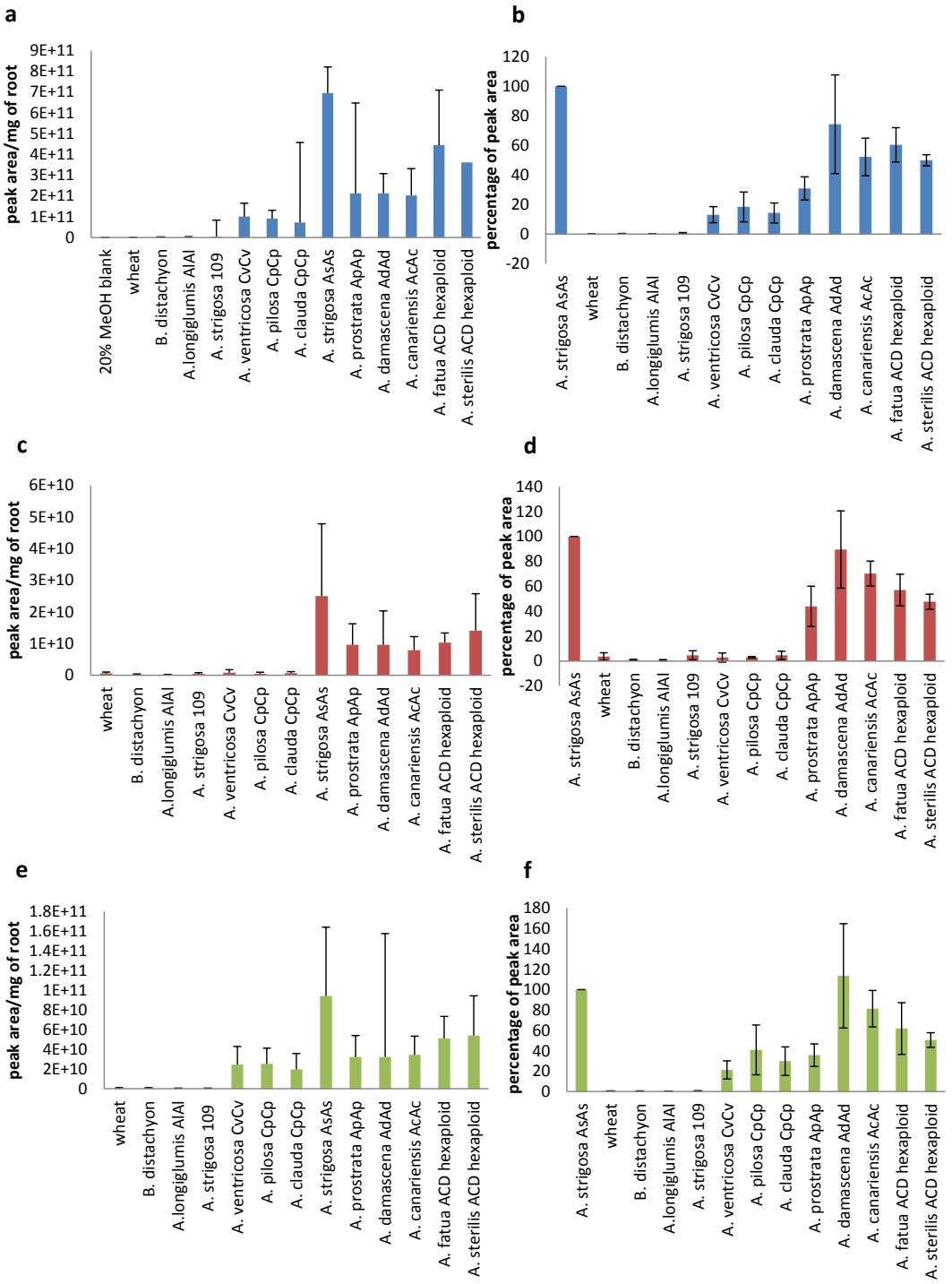


Figure 2.11: LC-MS of root extracts of *Avena spp.* a, c, e) The average amount of detectable avenacins A-1, A-2, and B-1 (measured in peak area from the LC-MS spectrum per mg of fresh root used to produce the root extracts). The error bars indicate standard deviations (four biological replicates).
b, d, f) The average relative abundance (percentage) of detectable avenacins A-1, A-2, and B-1 compared to that of *A. strigosa* S75 (as 100% abundance). The error bars indicate standard deviations (four biological replicates).

avenacin biosynthetic pathway within the genus *Avena*.

2.5.2. Variation in avenacin content within the genus *Avena*

The analysis of avenacin content carried out here is consistent with the previous report (Crombie and Crombie, 1986) that the bright-blue root fluorescence associated with avenacin A-1 is present in all *Avena* species surveyed with the exception of *A. longiglumis*. In contrast, such avenacin A-1-associated root fluorescence is absent in the other surveyed species within the *Aveninae*, wheat, and *B. distachyon*.

Metabolite analysis of root extracts confirmed that oat species with the root bright-blue fluorescence produce avenacins, while plants without the avenacin A-1 associated fluorescence are avenacin-deficient. The C genome oats, *A. clauda*, *A. ventricosa*, and *A. pilosa* produced lower quantities of avenacins than species with the A genomes. In conclusion, the avenacin content of the roots was dependent on the genotype of the surveyed species.

The higher intensity of avenacins detected in the roots of A genome oats compared to the C genome diploid species may be due to differences in regulation of the avenacin biosynthetic genes. On the other hand, the avenacin deficiency in the roots of *A. longiglumis* may be due to absence of the avenacin biosynthetic genes or presence of non-functional avenacin biosynthetic genes. Interestingly, *Sad1* expression has previously been detected in *A. longiglumis* root tips (Haralampidis et al., 2001). Further experiments investigating why oats with different genotypes produce different levels of avenacins will involve examination of the presence/absence of functional *Sad* genes and their expression patterns amongst the surveyed oat species.

2.5.3. Avenacin pathway relating to species phylogeny

Loskutov (2008) regarded *A. longiglumis* as possessing the ancestral A genome of the recently evolved A genome variants (Figure 2.2). Therefore, it was tempting to speculate that avenacin biosynthesis and the avenacin gene cluster emerged during the evolutionary trajectory of primitive Al-like genomes to more advanced As genomes (from *A. longiglumis* to *A. strigosa*) (Figure 2.2). However, the root-specific avenacin distribution that is present in most *Avena spp.* suggests that the biosynthesis of avenacin is common among most oats with either A genomes or C genomes (Figure 2.8 and 2.11) (except *A. longiglumis*). In addition, our phylogenetic studies of the oat phylogeny show that *A. longiglumis* is unlikely to

be the most ancient A genome oat.

The A genome oats and C genome oats diverged relatively soon after the split from the ancestral *Avenastrum* (Figure 2.12). Hence, the avenacin pathway may have arisen before the speciation of *Avena spp.* and the avenacin-deficient phenotype of *A. longiglumis* is likely to be due to loss of the avenacin biosynthetic genes (Figure 2.12).

According to previous investigations (Badaeva et al., 2010a; Loskutov, 2008), the most likely evolutionary pathway of *Avena spp.* followed a divergence of *A. macrostachya*, which possesses a CmCm genome, to the A and C genome diploid oat species. Although such inference is not fully in agreement with the phylogenetic analysis performed here, it is likely that *A. macrostachya* also has a functional and fully evolved avenacin biosynthetic pathway. It is also possible that avenacin biosynthesis may extend beyond *Avena spp.* but only by surveying more species in the *Pooideae* will the species boundary of avenacin biosynthesis be further delineated.

Overall, phylogenetic and metabolite analyses indicate that the avenacin biosynthetic pathway is present in most oat species but whether the *Sad* genes are clustered or not in oat genomes other than that of *A. strigosa* S75 is unknown. Inter-specific comparative studies of the avenacin gene cluster structure and genomic orientation of the *Sad* genes is not straight-forward due to a lack of sequence information about oat genomes. Nonetheless, conservation of the gene cluster amongst different oat species can be examined indirectly through assessment of the presence/absence of *Sad* gene homologues and their expression patterns. This is addressed in the next chapter.

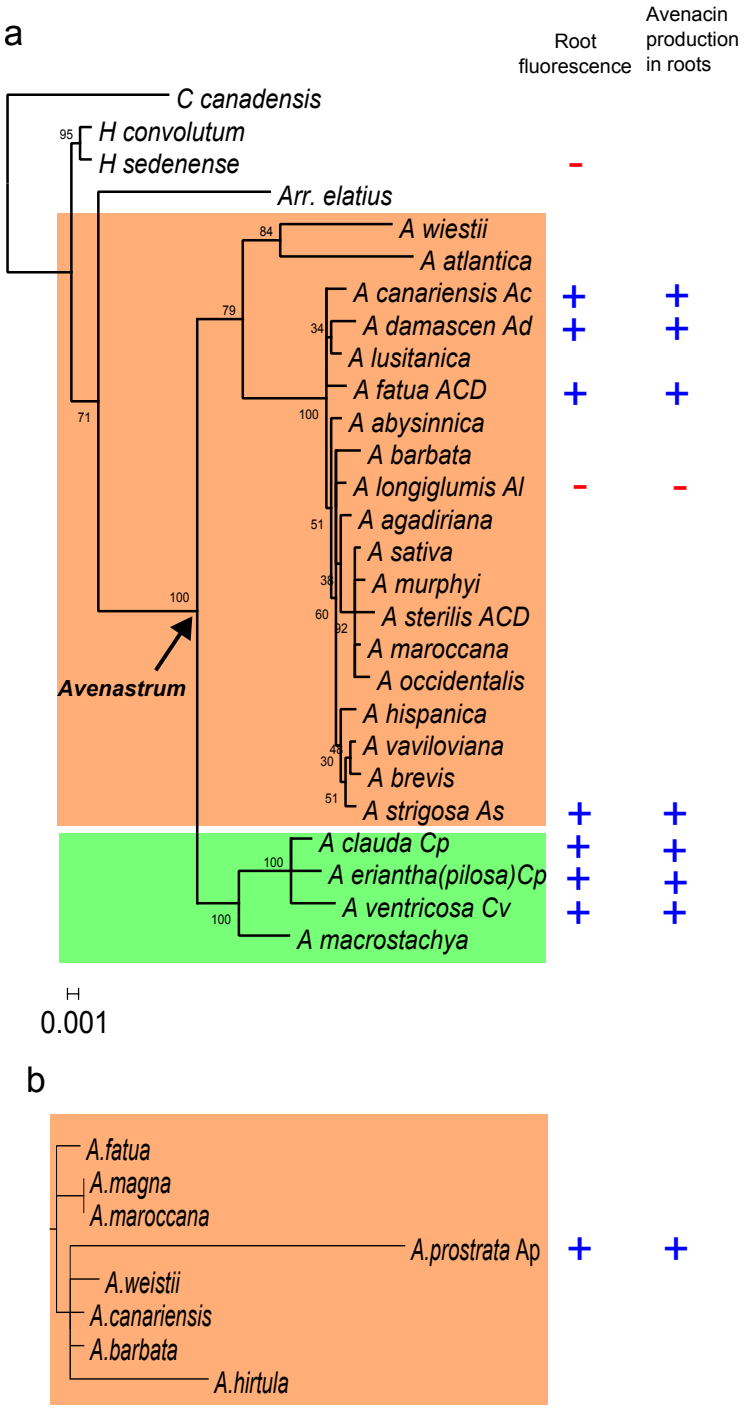


Figure 2.12: Avenacin biosynthetic ability superimposed onto the oat phylogeny. a) The RAxML TE tree with partitioning (Figure 2.6a), upon which root fluorescence and the ability to synthesise avenacins is indicated. b) the subclade within the A genome oats of the summarized RAxML ITS tree (Figure 2.5a) demonstrating the likely inter-specific relationship of *A. prostrata* with other A genome species. Its root fluorescence and the ability to synthesise avenacins is indicated. The A genome oats are highlighted in orange and the C genome oats in green. The node indicating the location of the ancestral *Avenastrum* is indicated by an arrow. Species that either produce or lack avenacins are indicated by (+) or (-) in the avenacin column (from previous publications, our preliminary root fluorescence screen and semi-quantitative analysis).

Chapter 3 - Conservation of avenacin biosynthetic genes within oats

3.1. Introduction

The previous chapter describes analysis of the root fluorescence and avenacin content within ten *Avena* species across the oat genealogy. It has been shown that all *Avena spp.* possess avenacin A-1-associated bright blue UV fluorescence in the roots with the exception of *A. longiglumis*. TLC and LC-MS confirmed that all oat species examined produced avenacins A-1, B-1 and A-2 but that *A. longiglumis* did not. Interestingly, the roots of seedlings of A genome oats contained around two-fold more of these avenacins compared to their C genome counterparts.
Here, the conservation of the avenacin biosynthetic (*Sad*) genes and their expression patterns across diverse oat accessiopns was investigated. Wheat and *B. distachyon* were also included in these experiments to gain a wider perspective across the *Pooideae*.

3.2. Materials and methods

3.2.1. Southern blot analysis

Total DNA was isolated from leaf tissues using cetyltrimethyl ammonium bromide (CTAB) solution followed by phenol chloroform extraction. Aliquots (15 μg) of total DNA were digested with 50-100U of Xba1 (Invitrogen^{TM}). The DNA digests were electrophoresed on 0.8%(w/v) agarose gels and transferred to Hybond N^+ membrane (GE Healthcare).
Radioactive probes for *Sad1, 2, 7, 9*, and *10* were produced from the full-length cDNA of the cloned *Sad* genes from *A. strigosa* accession S75 using the Invitrogen^{TM} random priming system, according to the manufacturer′s instruction. The *A. strigosa* S75 cycloartenol synthase (*AsCAS1*) full-length cDNA was used as a positive control. The ^{32}P dCTP labelled probes were hybridized with the DNA blots in Rapid-hyb buffer (GE Healthcare) according

to the manufacturer′s recommendation. After hybridisation, membranes were washed with 0.1% SSC, 1% SDS solution five times, each for 1 hour, at 55°C, 60°C, 61.5°C, 63°C, and 65°C. After each wash, the membrane was exposed to a phosphor-storage screen (GE Healthcare) for 16 hours and the image was developed in the Typhoon Scanner™ FLA9000 with the phosphor-storage screen setting.

3.2.2. Isolation of total RNA

Plant material

For each experiment, seeds of each species were germinated as described in section 2.2.2. Root tips (the terminal 5mm of the root) and leaf material were harvested from 5-day old seedlings. For each species, total RNA from approximately 50 mg of fresh material was extracted using the Qiagen RNAeasy™ kit; for larger amounts of starting material (>100 mg) total RNA was extracted using TRI-Reagent (Sigma). Extracted RNA was treated with Ambion®TURBO™ (Invitrogen) DNase to remove DNA. RNA concentrations were determined spectrophotometrically using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Scientific™).

3.2.3. Northern blot analysis

Total RNA (10 μg) from leaf and root tissues of seedlings was denatured in 2x RNA loading dye (New England Biolabs). The denatured total RNA (10 μg) was resolved on a 1 x MESA buffer (40 mM 3-(N-morpholino)propanesulfonic acid (MOPS), 10mM sodium acetate and 1mM EDTA (pH 8.3)) (Sigma)/1.2% agarose/0.25 M formaldehyde (37% formaldehyde solution (Sigma))/ 0.01 μg/ml ethidium bromide gel in 1 x MESA buffer at 60V for 3 hours. A single-strand RNA ladder (New England Biolabs) was run on the RNA gel alongside the denatured RNA samples as the size reference. Total RNA was transferred to a Hybond-N^+ nylon membrane (GE Healthcare Life Science) for northern blot analysis. Methylene blue staining (0.02% w/v methylene blue, 0.3M sodium acetate (pH 5.5)) of the blots was carried out to assess RNA transfer (Herrin and Schmidt, 1988). The 25S rRNA and 18S rRNA were used as size references for the hybridisation signals.

Radioactive probes were generated from the coding sequences of the full-length *Sad* genes via PCR reactions using plasmids containing the full-length cDNA of *AsCas1*, *Sad1* and *2*. Polymerase chain reaction (PCR) was carried out

according to the manufacturer′s instructions for the Phusion® High-Fidelity DNA polymerases (New England Biolabs) (2x 50ul reactions per probes) with 32P-dCTP. A mock PCR reaction in which the αP^{32}-dCTP was replaced with non-radioactive dCTP was included in each PCR reaction and the PCR product confirmed by 0.8% agarose gel electrophoresis in 1 x TAE buffer with ethidium bromide staining. The radioactively labelled probes were then purified with illustra MicroSpin G-50 Columns (GE Healthcare). Hybridisation of the radioactive probes with the RNA blots was carried out using PerfectHyb^{TM} Plus hybridisation Buffer (Sigma) under the manufacturer′s instructions. After hybridisation, the membrane was washed in 0.1 % SSC, 1%SDS solution five times, each for 1 hour, at 55°C, 60°C, 61.5°C, 63°C, and 65°C. After each wash, the membrane was exposed to a phosphor-storage screen (GE Healthcare Life Science) for 16 hours and the image was developed in the Typhoon Scanner^{TM} FLA9000 with the phosphor-storage screen setting.

3.2.4. Reverse transcription polymerase chain reaction (RT-PCR) analysis

Reverse transcription was performed with 2 μg of total RNA using the SuperScript® First-Strand Synthesis kit (Invitrogen), according to the manufacturer′s instructions, for first strand synthesis of transcripts with high GC content using Oligo$(dT)_{15}$ primers. Isolation of the full-length cDNA were carried out with PCR using Phusion® High-Fidelity DNA polymerases (New England Biolabs) (Table 3.1). The actin gene was used as a positive control for cDNA quality. The primers were designed using the cloned *Sad* genes of *A. strigosa* S75. The PCR products were resolved on an 8% agarose electrophoresis in 1 x TAE buffer and products of interest were extracted from the gel.

3.3. Results

3.3.1. Southern blot analysis to establish the presence/absence of *Sad* genes among *Avena spp.*

The results of the Southern blot analysis are shown in Figure 3.1. A total of 62 *Xba1* restriction sites are found in the sequences of the BAC contigs encompassing the five characterized *Sad* genes in *A. strigosa* S75 (Mugford et al., 2013). The anticipated lengths of *Xba1*-digested fragments derived from these *A. strigosa* BAC contig sequences are shown in Table 3.2.
Southern blot analysis of *A. strigosa* S75 genomic DNA probed with *Sad1*

Gene target	Forward primer ($5' \rightarrow 3'$)	Reverse primer ($5' \rightarrow 3'$)	Product length (bp)	Annealing temperature (°C)
Sad1	Sad1-1-5 ATGTGGAGGCTAA CAATAG	AMY_END_R1 TGATGACATCGG TAGGAA	2274	53
Sad2	CypA_ATG_F ATGGACATGA CAATTTGC	AsCypAedRns GTTTGCAGGCAT ACGACATCTCT	1473	54
Sad7	S7_20F GTGGTGCTGC TGCTAGTGAC	Sad7′q3 GTGGTGCTGC TGCTAGTGAC	1443	54
Sad9	Sad9-1-5 ATGGGGCATG TCCACACTAC	Sad9ENDRs CAATGATAGA TCGAAATCCCAA	1046	52.5
Actin	actin-5′ CCCCGTCTGC GACAATGGTA	actin-3′ TCCTCTCGCT GTACGCCAGT	470	50

Table 3.1: Primers for amplification of full-length coding sequences of *Sad* genes from cDNA

Gene	Expected restriction fragment length (bp)
Sad1	15051, 7663
Sad2	3171, 7915
Sad7	20861
Sad9	15214
Sad10	7632

Table 3.2: Anticipated lengths of *Xba1*-digested fragments of *Sad* genes in the *A. strigosa* S75 avenacin gene cluster.

revealed the two anticipated hybridisation bands with *Sad1* probes of length 7.7 kb and 15 kb (Figure 3.1). The other A genome oats exhibited one or two strong hybridisation signals of varying sizes with the *Sad1* probe. The C genome oats shared a very similar pattern to each other, with three clearly hybridising bands. Multiple hybridisation bands were observed for *A. fatua* and *A. sterilis*. Hybridisation was not detected for wheat while *B. distachyon* exhibited a very weak band when probed with *Sad1*. The absence of a signal for *Sad1* in wheat is unlikely to be due to low loading levels since *Xba1*-digested genomic DNA was clearly detectable by ethidium bromide staining (Figure 3.1g).

When probed with *Sad2* (Figure 3.1b), *A. strigosa* S75 exhibited the two expected labelled fragments of around 3.1 and 7.9 kb. A weaker band of over 10 kb was also detected, suggesting the presence of additional *Sad2*-like sequences. The other A genome oat species exhibited one or two strong hybridisation signals of variable size, while the C genome species all showed a conserved pattern with three fairly weak signals. *A. sterilis* and *A. fatua* both had several bands. The hybridisation

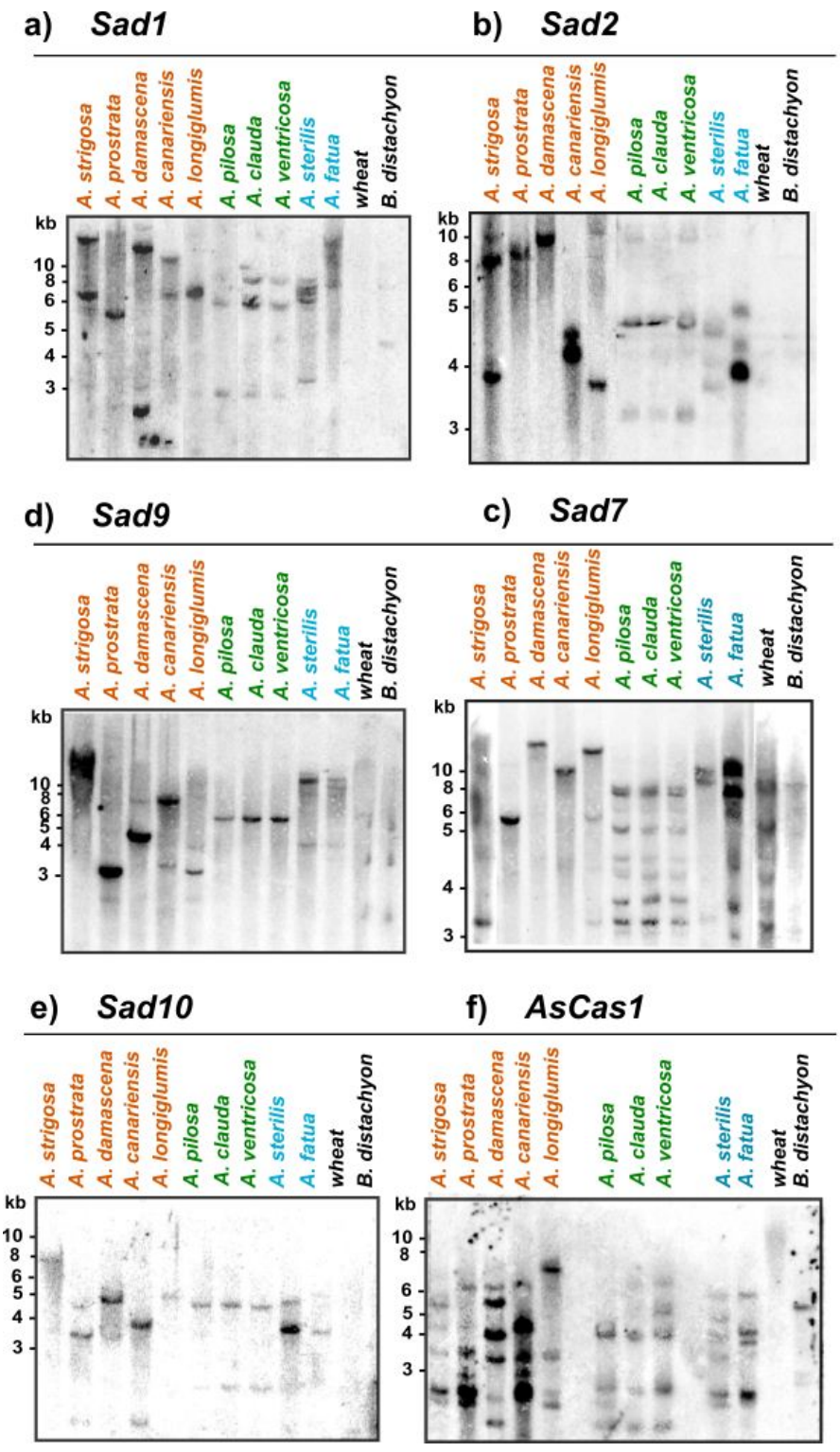


Figure 3.1: Southern blot analysis a-f) of *Xba1*-digested genomic DNA showing hybridisation to radioactively labelled probes specific for *Sad1, 2, 7, 9, 10* and *AsCas1* (positive control). The blots washed under the optimal stringency was shown. The image of the ethidium bromide-stained gel showing *Xba1*-digested genomic DNA from different species (0.8% agarose gel run in 1 x TAE buffer) is shown in the Appendix Figure 3.1. A genome oats are highlighted in organe, C genome oats in green and the hexaploid species in blue.

patterns of the C genome species are highly similar while those of the A genome oats are diverse. Wheat and *B. distachyon* did not exhibit any *Sad2* hybridisation. The *Xba1*-digested restriction fragment expected for the *Sad7* probe in *A.strigosa* S75 was anticipated to be approximately 20 kb. A band of this size was beyond the size range of detection. A smaller hybridizing band of unknown identity around 3.5 kb was, instead, detected (Figure 3.1c). While the A genome oats each exhibited a single strongly hybridising band of various sizes, approximately eight bands were detected in the C genome species and six in *A. fatua*, wheat and *B. distachyon* suggesting the presence of multiple sequences that are highly similar to *Sad7* in the genomes of these species.

When probed with *Sad9* (Figure 3.1d), *A. strigosa* S75 exhibited the predicted labelling pattern of one band of around 15 kb. The A and C genome species and *A. sterilis* each gave one strong hybridisation signal while there were two weaker hybridisation signals in *A. fatua*. Wheat and *B. distachyon* appeared to exhibit three weak hybridisations signals.

The *Sad10* probe gave a band of around 7.6 kb as predicted (Table3.2, Figure 3.1e). Clear bands were also detected with this probe in most oat accessions, and a weak hybridisation signal in *A. longiglumis*. No clear hybridisation signal was observed with wheat and *B. distachyon* (Figure 3.1e).

A probe for *A. strigosa* cycloartenol synthase gene *AsCas1* was also included in these experiments. The genes for cycloartenol synthase, encoding the first enzyme for primary sterol biosynthesis, are highly conserved across higher plants. Hybridisation with the *AsCas1* probe gave signals for all species tested except wheat (Figure 3.1f).

In summary, all *Avena spp.* investigated contain sequences that are highly similar to the five cloned *Sad* genes of *A. strigosa*. In addition, weak signals were detected in *B. distachyon* with the *Sad1, 7* and *9* probes suggestive of the presence of distantly related sequences, while clear hybridisation was observed in wheat with the *Sad7* and *9* probes.

3.3.2. Expression analysis of *Sad* gene homologues

To determine whether the *Sad* gene homologues present in the *Avena spp.* studied have root-specific expression patterns as in *A. strigosa* S75, expression profiles in 5-day old seedlings were first inspected by RT-PCR and then by northern blot analysis. RT-PCR relies on a high level of sequence similarity between the sequences of the *Sad* gene homologues in other species and the primers used for transcript amplification. In contrast, northern blot analysis allows hybridisation of the cDNA-derived probes to homologous targets at a range of different stringencies

and so is more robust and flexible.

RT-PCR analysis revealed the presence of transcripts of *Sad1, 2, 7*, and *9* in the roots but not the leaves of seedlings of *A. strigosa* S75, as excepted (Figure 3.2). Amplification of *Sad10* cDNA were not successful using the available primers and thus *Sad10* was excluded from the RT-PCR analysis. This is consistent with previous findings (Haralampidis et al., 2001; Mugford et al., 2013, 2009; Qi et al., 2004, 2006; Wegel et al., 2009). Transcripts of *Sad1, 2* and *7* homologues were detected by RT-PCR in the roots of all *Avena* species including *A. longiglumis*, which is avenacin-deficient (Figure 3.2). Transcripts of *Sad* gene homologues was not observed in young root or leaf tissues of wheat and *B. distachyon* (Figure 3.2). Although *Sad*-like sequences were detected in the genomes of wheat and *B. distachyon* by southern blot analysis for *Sad1, 7* and *9*, no transcripts corresponding to homologures of these genes were detected in the RT-PCR analysis. This may be because either the sequences detected by southern blot analysis lack the primer sites for RT-PCR amplification or because the corresponding genes are not expressed in the leaves and roots of the seedlings. A barely detectable RT-PCR product of similar size to the *Sad2* transcript was detected with RNA from the leaves of *A. sterilis* seedlings, suggesting that a *Sad2* homologue or closely related gene is expressed in the leaves of this species. Of note, RT-PCR detected transcripts for *Sad1, 2,* and *7* gene homologues in both root and leaf RNA in the C genome diploid oat seedlings (Figure 3.2). Sequence information of the RT-PCR detected transcripts are obtained and are analysed in next Chapter.

RT-PCR analysis indicated only low levels of amplification of transcripts using primers designed to detect *Sad9* with RNA from the roots of seedlings of *A. prostrata, A. canariensis, A. pilosa, A. clauda* and *A. ventricosa*. This may be due to low levels of expression of *Sad9* homologues or alternatively may indicate that the *Sad9*-like sequences in these species are highly divergent from that of *A. strigosa* S75. No *Sad9* transcripts were detected in *A. longiglumis* (Figure 3.2), which may be due either to sequence divergence of *Sad9* in the *A. longiglumis* genome or to a lack of expression.

Northern blot analysis was performed for the first and second genes in the pathway (*Sad1* and *Sad2*) to reinforce the RT-PCR experiments (Figure 3.3). The A genome species all gave a single strong hybridisation signal for root RNA when probed with *Sad1* or *2*, consistent with the RT-PCR analysis data. A weak *Sad1* hybridisation signal was observed in the RNA from wheat leaves, suggesting the presence of a leaf expressed gene that is distantly related to *Sad1*. In contrast, hybridisation was observed for both root and leaf RNA of the C genome species

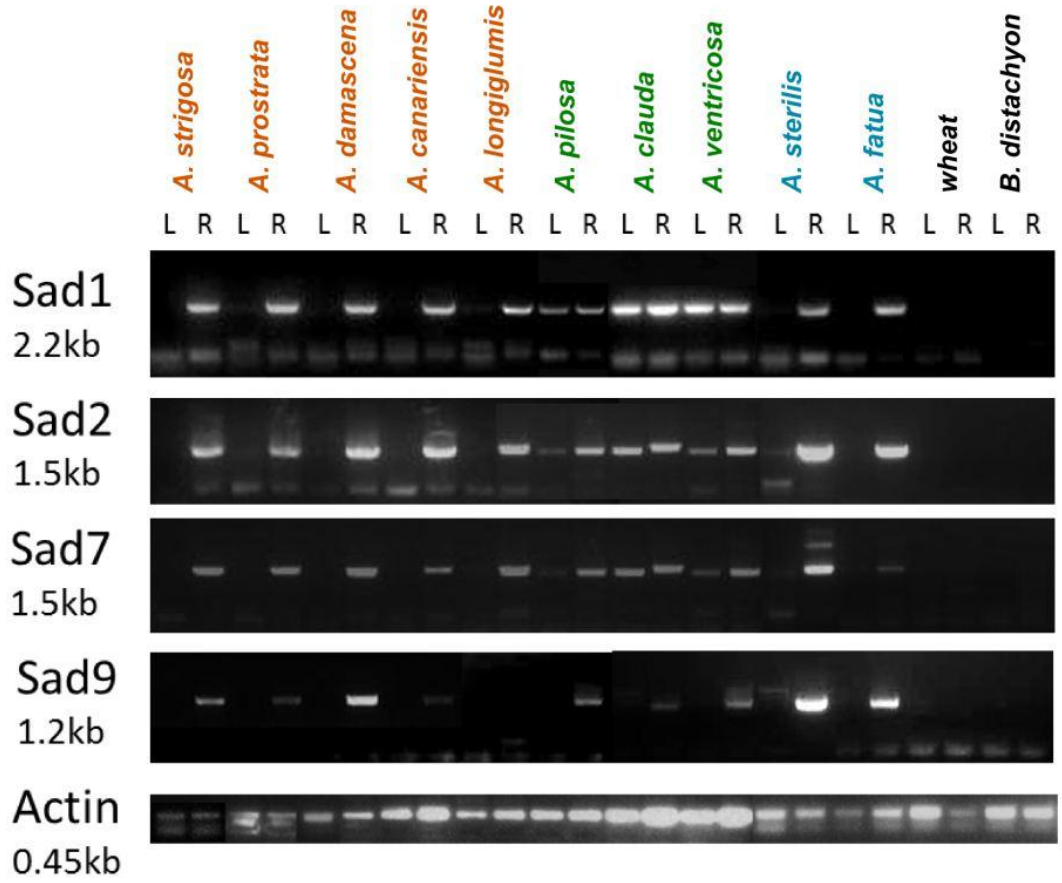


Figure 3.2: RT-PCR analysis of transcripts of *Sad1, 2, 7* and *9* homologues. RT-PCR was performed using cDNA derived from the roots and leaves of five-day old seedlings. *A. strigosa* S75 is included as a positive control; negative controls are wheat and *B. distachyon*. cDNA from leaf (L) and root (R) material are indicated. A genome oats are highlighted in orange, C genome oats in green and hexaploid species in blue. RT-PCR of the actin gene was performed as a positive control of cDNA synthesis. The expected sizes of RT-PCR products of *Sad1, 2, 7, 9* and the actin gene were indicated.

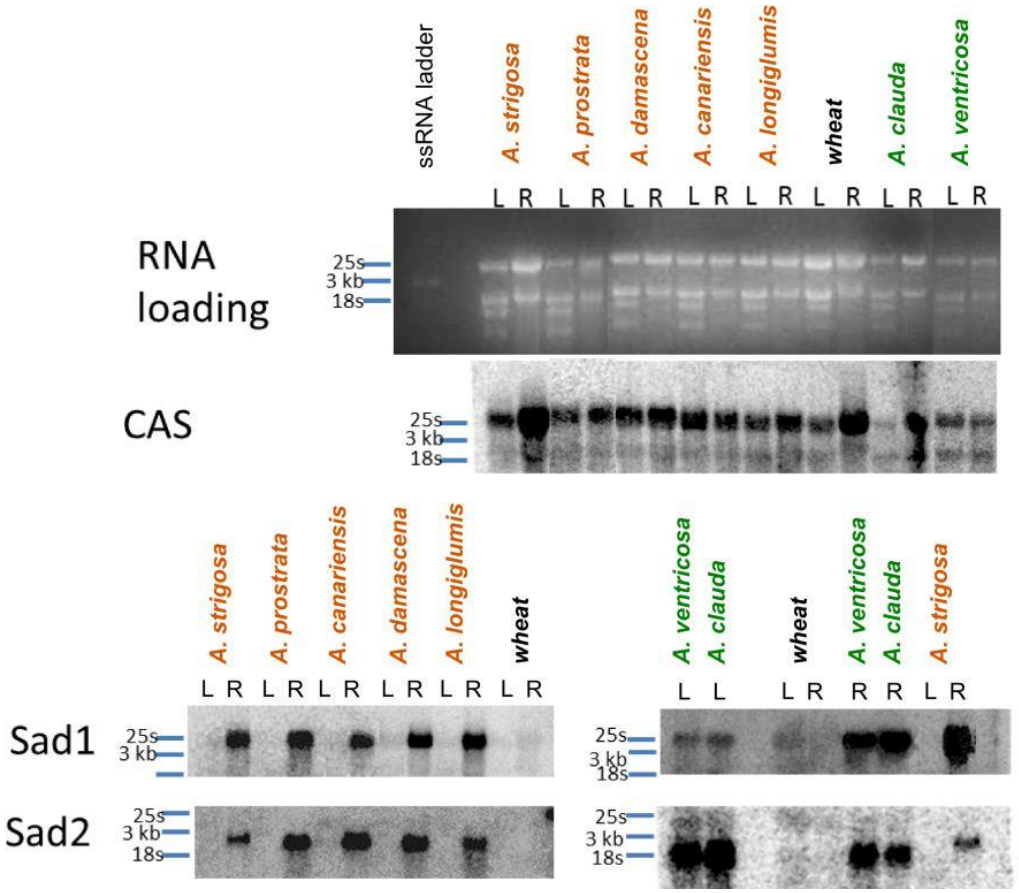


Figure 3.3: Northern analysis Blots of total RNA from root and leaves of 5-day old seedlings of *Avena spp.* and wheat were hybridised with probes for *Sad1*, *Sad2* and *AsCas1* (positive control). Labels: RNA from leaf (L) and root (R) material. A genome oats were highlighted in orange and C genome oats in green. RNA loading was monitored with ethidium bromide.

with the (*Sad1* and *2*) probes. Of note, the *Sad2* root transcripts of *A. strigosa* appear to be larger than those of the C genome oats.
The transcripts detected in the A genome species with the *Sad2* probe have a size of approximately 3 kb. In contrast, the leaf and root transcripts detected in the C genome oats using this probe have a size of approximately 1.9 kb. These different transcript sizes may be due to differences in the lengths of untranscribed regions or post transcriptional modifications such as polyadenylation.

3.4. Discussion

3.4.1. *Sad* gene distributions in oat genomes

Southern blot analysis showed that the genome of all A genome oats, including the avenacin deficient *A. longiglumis*, contain the five *Sad* genes investigated. Because *A. longiglumis* is not able to synthesise avenacin, it is possible that one or more of the *Sad* gene sequences detected in the southern analysis are pseudogenised or transcriptionally repressed. The C genome oats all exhibit weak hybridisation with the probes in the southern analysis, suggesting that the *Sad* gene sequences of the C genome species are divergent from those of the A genome oats. The diverse restriction patterns of the *Sad* genes among the A genome oats in the southern blots may be suggestive of a diverse avenacin gene cluster organisation or broader genomic distribution of the *Sad* genes.

3.4.2. Expression analysis of *Sad* gene homologues

RT-PCR and northern blot analyses collectively showed that *Sad1, 2, 7* and *9* of A genome oats are expressed only in the roots but not in the leaves of five-day old seedlings. Importantly, *A. longiglumis* was found to contain full-length transcripts of *Sad1, 2* and *7*, the key genes in avenacin biosynthesis, in roots. These results suggest that at least three of the *A. longiglumis Sad* genes are likely to be functional unless these full-length transcripts of *Sad1, 2* and *7* are not translated into proteins in *A. longiglumis* roots. In contrast, RT-PCR did not detect expression of *Sad9* homologues in the roots of *A. longiglumis*, suggesting that *Sad9* may be non-functional. If this is the case, *A. longiglumis* should behave like the *sad9* mutant, accumulating the intermediates des-methyl avenacin A-1 (DMA) (Mugford et al., 2013). However, LC-MS on *A. longiglumis* root extracts performed in Chapter 2 did not detect any DMA nor Des-acyl avenacin A-1 (DAA), which is an earlier pathway intermediate (upstream of the step catalysed by SAD7) (Mugford et al., 2013, 2009). Therefore, the presence of

the *Sad1, 2* and *7* transcripts is unlikely to contribute to any avenacin synthesis at all. In the future, western analysis probing the presence/absence of the *Sad* gene products in *A. longiglumis* roots will give more insights into the key factor of its avencin deficiency.

Although avenacin gene expression is root specific to the A genome oats, *Sad1, 2* and *7* transcripts have been detected in both root and leaf RNA of the C genome oats. In *A. clauda*, two *Sad2* transcripts isoforms were found in young seedlings roots. The presence of at least two copies of *Sad* gene homologues in the C genome oats may account for the leaf expression of *Sad 1, 2* and *7*. Alternatively, the avenacin genes may instead be regulated in a distinct manner in the C genome oats, expressing in both root and leaf tissues in young seedlings. However, this latter hypothesis is unlikely as the expression of *Sad9*, which is also in the gene cluster, shows root-specific expression in the C genome oats. Further sequence analysis of these transcripts from leaves and roots to dissect their evolutionary origin is performed in next Chapter.

Chapter 4 - Analysis of the oat *Sad* gene homologues

4.1. Introduction

The work described in the previous chapter involved investigation of the presence/absence of homologues of avenacin biosynthetic genes and gene expression analysis within ten *Avena* species. Southern blot analysis showed that all oat species examined contained homologues of the five characterized *Sad* genes (*Sad1, 2, 7, 9* and *10*). RT-PCR and northern blot analysis revealed that homologues of *Sad1, 2, 7* and *9* were expressed exclusively in the roots of A genome oats. *A. longiglumis* was found to have root-specific expression homologues of *Sad1, 2* and *7* but not *Sad9*. In contrast, transcripts of homologues of *Sad1, 2* and *7* were detected in both leaves and roots in the C genome oat species. To further develop our understanding of the diversity of the *Sad* gene homologues amongst different species, the DNA sequences of the genomic sequences and the transcripts of the *Sad* gene homologues detected in the RT-PCR analysis of the different oat species were determined and analysed in this chapter.

4.2. Materials and methods

4.2.1. PCR amplification of *Sad* gene genomic fragments

Genomic DNA of the selected *Avena spp.* (Table 2.1) was extracted from 10g of plant material as described in 3.2.1. PCR was performed using a combination of primers for *Sad1, 2, 9* and *10* (Table 4.1) designed using the *Sad* gene sequences of *A. strigosa* S75. PCR reactions were carried out using the Phusion® High-Fidelity DNA polymerases (New England Biolabs) to amplify the designated parts of the *Sad* genes. The PCR products of 50μl reactions were confirmed by 0.8% Tris-acetate buffer with EDTA (pH 8.0) (TAE) agorose gel electrophoresis with ethidium bromide staining. PCR products were purified using QIAquick PCR purification kit (Qiagen) and sent to the Genome Enterprise Ltd or Eurofins MWG Operon for direct sequencing.

Primers used for PCR amplification of *Sad* genes				
Primer name	Gene target	Primer sequence(5′-3′)	Primer pair	Annealing temp. (°C)
AsCAS_F1	*Cas1*	TAATGTGGCGGCTGAAGATC	AsCAS_F1	60
AsCAS_R1	*Cas1*	ACTACTTGCCCGCAGCCAA	AsCAS_R1	
actin-5′	Actin	CCCCGTCTGCGACAATGGTA	actin-5′	52
actin-3′	Actin	TCCTCTCGCTGTACGCCAGT	actin-3′	
Sad1-1-5	*Sad1*	ATGTGGAGGCTAACAATAG	Sad1-1-5	53
AMY26R	*Sad1*	ATCGTCTGCTTGTAGAGAGGA	AMY26R	
AMY18F	*Sad1*	GTGGCTCATCACATTGATCACA	AMY18F	62
AMY610R	*Sad1*	CATACCGACAACCATATTTTTCCCCA	AMY610R	
AMY02F	*Sad1*	GTTGGGGAAAAATGGTTC	AMY02F	52
AMY24R	*Sad1*	GTCCCAATTAATGTTGCAGTAAG	AMY24R	
AMY109R	*Sad1*	TCTATACCAACCTGTGCCTTCATTCC	AMY02F AMY109R	52
S1_2767F	*Sad1*	TGGTGTTTTACCCGGTTGAT	S1_2767F AMY109R	58
AMY03F	*Sad1*	TTTTCTTCCGATTCACCC	AMY03F	60
S1_3507R	*Sad1*	CTGTGCCTTCATTCCATCCT	S1_3507R	
AMY05F	*Sad1*	TGGATGTCATAGCTGGGA	AMY05F	53
AMY23R	*Sad1*	TAGTCCACGACAATGTTCCGA	AMY23R	
AMY12F	*Sad1*	CTCAACCCTTCTGAGAGTTTT	AMY12F	56
AMY21R	*Sad1*	ATATTTGACAAACCTGTCCAGCAT	AMY21R	
S1_5950F	*Sad1*	ATATTTGACAAACCTGTCCAGCAT	S1_5950F AMY21R	56
S1_5906R	*Sad1*	AGTCTTCACCCCATCCACCT	AMY12F S1_5906R	56
AMY15F	*Sad1*	GGCGGAGGATGGAATGAAGGCA	AMY15F	63
AMY17R	*Sad1*	AGCCTTTTGATCTGTGGCGATA	AMY17R	
S1_2079F	*Sad1*	GCAATACCACAGTGGGGAAA	S1_2079F S1_3507R	54
S1_3271F	*Sad1*	TATCCATTATGACGACGAATCAACC	S1_3271F	54
S1_3926R	*Sad1*	TCCATCATATACCTGCGATTGTATAC	S1_3926R	
AMY19F	*Sad1*	TGGGCAATGTTGGCTTTAATTT	AMY19F	60
S1_d212R	*Sad1*	AAAACAACGATTAAACCGCG	S1_d212R	
AMY32R	*Sad1*	CTGTAAGTAAGAAGCTAGATGGAG	AMY19F AMY32R	58
S2_u268F	*Sad2*	TTCGATAACAATCACGCATCA	S2_u268F	56
S2_366R	*Sad2*	CCATTCTCTTTGCCGAACAT	S2_366R	
S2_175F	*Sad2*	GACCCTGAAGCTGCAATGAT	S2_175F	60
S21851R	*Sad2*	CTTGTGTGCTTTCCAGCAAA	S2S2_1851R	
Sad2q′3	*Sad2*	ATCTCGGACCTCACTTCCAA	S2_175F Sad2q′3	56

continued on next page

continued from previous page				
Sad2q′5	*Sad2*	TCGACAGGAAGTGGAGG	Sad2q′5 S2_1861R	58
Isu441gF1	*Sad2*	ACGAGGGTGAAGTCGATCTGAAA-CAAGAG	Isu441gF1	63
AsCypA-edRns	*Sad2*	GTTTGCAGGCATACGACATCTCT	AsCypA-edRns	
S1_2074R	*Sad2*	CACCTCCCTCTCTTGTCTGC	Isu441gF1 S2_2074R	58
S7_u82F	*Sad7*	TCGTGGCAGAAATAGGGTTT	S7_u82F	65
S7_490R	*Sad7*	TCCACTTCTGGAGGAACACC	S7_490R	
AsATF04	*Sad7*	GCTCAACAGCACCGTCACCG	AsATF04	58
Sad7q′3	*Sad7*	GATCCATCTTCGGACCATGT		
Sad7_771F	*Sad7*	GACGGATTCGAAGAACAAGC	Sad7_771F	58
Sad7_-1925R	*Sad7*	GAGTAAGTCATGGTCGCCGT	Sad7_1925R	
S9_u183F	*Sad9*	CCCAACAAGATACACGTCCA	S9_u183	55
S9_437R	*Sad9*	CCTTCTTGGCTGTCATCTCC	S9_437R	
Sad9-1-5	*Sad9*	CCTTCTTGGCTGTCATCTCC	Sad9-1-5	58
S9_941R	*Sad9*	ACCTTGATGGCATCTTCGTC	S9_941R	
S9_726F	*Sad9*	GGGGATGCGTTTCAGTACAT	S9_726F	55
S9_d190R	*Sad9*	TGCCATATTTGCTTCAAGACC	S9_d190R	
Sad10-3-5	*Sad10*	GTGGGAGCACGTCAGCGAC	Sad10-3-5	51
S10_1323R	*Sad10*	CTCTCCAACTCCTCCCTCCT	S10_1323R	
AsGTF03	*Sad10*	ACCAAGACATGCTCTTGTCTC	AsGTF03	53
S10_d133R	*Sad10*	CCAGCAGGGCCATAGTATTT	S10_d133R	

Table 4.1: Table of primer pairs used for *Sad* gene amplification in the PCR analysis

Reads generated from the direct sequencing results were assembled or concatenated to give 'full-length' *Sad* genes manually using BioEdit (Hall, 1999), using the *A. strigosa* S75 *Sad* gene sequences as references.

4.2.2. Retrieval of *Sad* gene coding sequences from RT-PCR

The RT-PCR products were extracted from 0.8% agarose gels using the Promega Wizard® SV Gel and PCR clean up system and then cloned into One Shot® Top10 cells (according to the protocol of the Zero Blunt® Topo® Cloning kit) for sequencing. At least three colonies per RT-PCR product were picked and the plasmids were purified using the QIAquick miniprep kit (Qiagen). The inserted PCR products were verified to be *Sad* genes by PCR amplification using the primers pairs used in the RT-PCR experiments (Table 3.1). These plasmids were then sent to the Genome Enterprise Ltd or Eurofins MWG Operon for direct sequencing.

4.2.3. Reconstruction and sequence comparison of coding sequences and translated amino acid sequences of *Sad* gene homologues

Consensus coding sequences of *Sad* genes of each species were generated from the direct sequencing results of at least three clones using the Vector NTI® ContigExpress software (Life technologies™). The consensus sequences were then annotated in Geneseqer (Brendel et al., 2004) using the amino acid sequences of the *A. strigosa Sad* genes as the reference sequence templates. The annotated coding sequences were then translated to amino acid sequences. These were then aligned using MUSCLE 3.6 (Edgar, 2004). The codon alignments were generated in Pal2Nal (Suyama et al., 2006) using the amino acid alignment as a reference. Non-synonymous differences between the reconstructed coding sequences and the *A. strigosa Sad* gene reference sequences were identified from the sequence alignment.

4.2.4. Phylogenetic studies of *Sad* genes

Phylogenetic trees were estimated from both reconstructed coding and amino acid sequence alignments of the RT-PCR products as well as the reconstructed gene sequences obtained from PCR amplification of *Sad* gene fragments from genomic DNA. Both amino acid and codon alignments were input into RAxML

7.0.3 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001), for phylogenetic tree construction with substitution models specified in the ProtTest and FindModel servers (summarised in 2.2.1). The resulting phylogenetic trees for *Sad1, 2* and *7* were rooted using published amino acid and coding sequences of *Sad1, 2* and *7* from *A. strigosa* S75 as outgroup sequences. The trees were then compared to the oat species phylogeny generated in Chapter 2.

4.2.5. Three-dimensional modelling of *Sad1*

The three-dimensional protein structure of the translated amino acid sequences for *Sad1* root transcripts of each *Avena spp.* obtained from RT-PCR were modelled using I-TASSER 2.0.1 (Zhang, 2008) server. The protein models were then compared and aligned using the UCSF Chimera software (Pettersen et al., 2004). The non-synonymous differences between the sequences of the A and C genome oats identified in the previous multiple sequence alignments were mapped to the structural alignments and the protein models.

4.3. Results

4.3.1. Retrieval of the genomic sequences of the *Sad* gene homologues

The genomic regions of *Sad1* homologues from the various *Avena spp.* were successfully amplified by PCR to give a total of 12 contigs, except for a 500 bp region spanning intron 2 that lacked available primer combinations yielding a single PCR product (Figure 4.1). Complete amplification of the genomic sequences of *Sad2* and *Sad7* homologues were obtained by PCR using primer pairs according to Figure 4.1. PCR amplification for *Sad9* and *10* yielded multiple products, which could not be purified. Thus genomic sequence information were not determined for *Sad9* and *10*. In all PCR amplications of the *Sad1, 2* and *7* homologues from genomic DNA, the resulting single PCR products of each PCR reaction were of the predicted sizes of the reference *A. strigosa* S75 sequence. These PCR products were purified and sent for direct sequencing. The sequencing results were assembled into one or more contigs and subsequently into 'full-length' genes, using the reference *A. strigosa* S75 *Sad* gene sequences as templates.

The assembled contigs for *Sad1, 2* and *7* gene homologues were subsequently analysed alongside the sequencing data obtained from RT-PCR (see Subsection 4.3.2).

Summary of *Sad1* trees						
Phylogenetic software	Alignment type	Number of sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	24	373 aa	-2405.67	1.59	JTT + Γ
	codon	24	1119 bp	-4135.67	0.85	GTR + Γ
MrBayes	amino acid	24	373 aa	-2086.95	0.42	JTT + Γ
	codon	24	1119 bp	-3330.53	0.30	GTR + Γ
Summary of *Sad2* trees						
RAxML	amino acid	25	366 aa	-2376.54	1.38	JTT + Γ
	codon	25	1098 bp	-3940.06	1.08	GTR + Γ
MrBayes	amino acid	25	366 aa	-1874.66	0.45	JTT + Γ
	codon	25	1098 bp	-2950.17	0.39	GTR + Γ
Summary of *Sad7* trees						
RAxML	amino acid	24	326 aa	-3271.26	1.79	WAG + Γ
	codon	24	978 bp	-5931.41	0.90	GTR + Γ
MrBayes	amino acid	24	326 aa	-2585.39	1.14	WAG + Γ
	codon	24	978 bp	-5979.36	0.51	GTR + Γ

Table 4.2: Summary of the phylogenetic analyses of the *Sad* genes carried out in RAxML 7.0.3 (Stamatakis, 2006) and MrBayes 3.2.1. (Huelsenbeck and Ronquist, 2001)

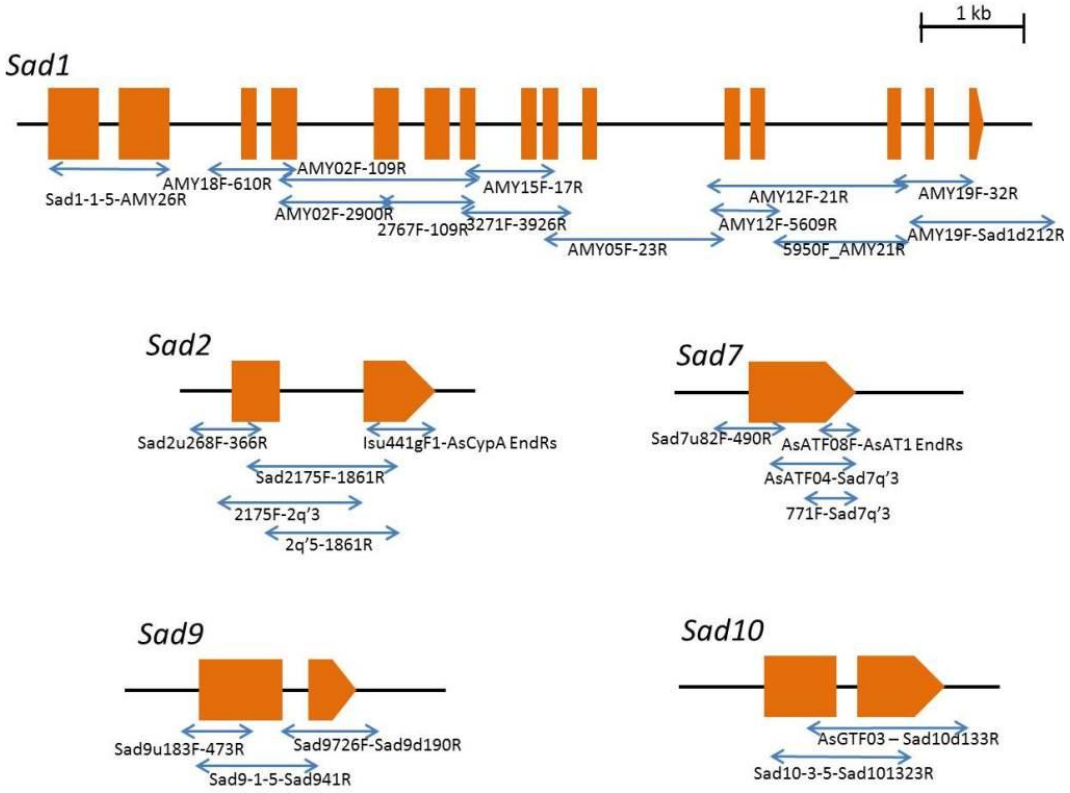


Figure 4.1: Diagram showing the strategy for sequencing of *Sad* gene homologues in other oat species. Orange arrows - exons, blue arrows - PCR products generated using the primers specified in Table 4.1.

4.3.2. Analysis of the *Sad* gene transcript sequences

To further validate the identities of the transcripts detected for homologues of *Sad1, 2, 7* and *9* in the RT-PCR analysis discussed in Chapter 3, sequence information was obtained. The sequences of the cloned *Sad* gene homologue transcripts were annotated and their coding sequences reconstructed using the reference amino acid sequence from *A. strigosa* S75. These reconstructed coding sequences were aligned and differences at the amino acid level were identified (Appendix Alignment 4.1 - 4.4). The non-synonymous differences within *Sad1, 2, 7*, and *9* are listed in Appendix Table 3.1. Finally, phylogenetic analyses (summarised in Table 4.2) of the retrieved genomic sequences (Subsection 4.3.1) as well as the transcripts of *Sad1, 2,* and *7* homologues were carried out and later compared to the inferred oat phylogeny (Figure 2.6 and 2.7) to trace the origins of these *Sad* gene homologues.

4.3.3. Analysis of oat *Sad1* homologues

Oxidosqualene cyclases (OSC) are enzymes that generate sterol/triterpene scaffolds. *Sad1* encodes an OSC that catalyses the formation of the triterpene scaffold β-amyrin from 2,3-oxidosqualene (Figure 1.2). The product specificity of OSCs is likely to be achieved by substrate folding into an appropriate conformation and subsequent stabilization of the cyclic intermediates by motifs known as QW motifs (Thoma et al., 2004; Wendt et al., 1997). OSC enzymes also have a conserved DCTAE motif buried in the cyclisation domain that possesses a C-D-C catalytic triad, of which the D residue initiates the cyclisation reaction (Thoma et al., 2004; Wendt et al., 1997). The GYN motif is a feature of OSCs that synthsise sterols and is present in cycloartenol and lanosterol synthases. This motif is substituted by VYD in the corresponding positions in *A. strigosa* β-amyrin synthase 1 (SAD1) (Inagaki et al., 2011).
Sequence comparisons amongst oat *Sad1* homologues (Appendix Alignment 4.1) show that the key residues for substrate entry (T267, Y271, and I561), those for enforcing ring conformation (Y126, W264, F266, L480, Y540, W614, F733 and F739) and the cyclisation domain (D491, C492 and C570) are all conserved (Figure 4.2) (Racolta et al., 2012; Thoma et al., 2004; Wendt et al., 1997). There are a total of 34 non-synonymous (NS) sites detected amongst the oat *Sad1* homologue sequences. More than half of the non-synonymous differences (20 out of 34 NS sites) detected in the alignment were between the C genome oat transcripts and the rest of the sequences (Appendix Table 4.1). Amongst the 34 NS differences, 27 of them were accounted by single-base changes in the codon. Of note, two of these

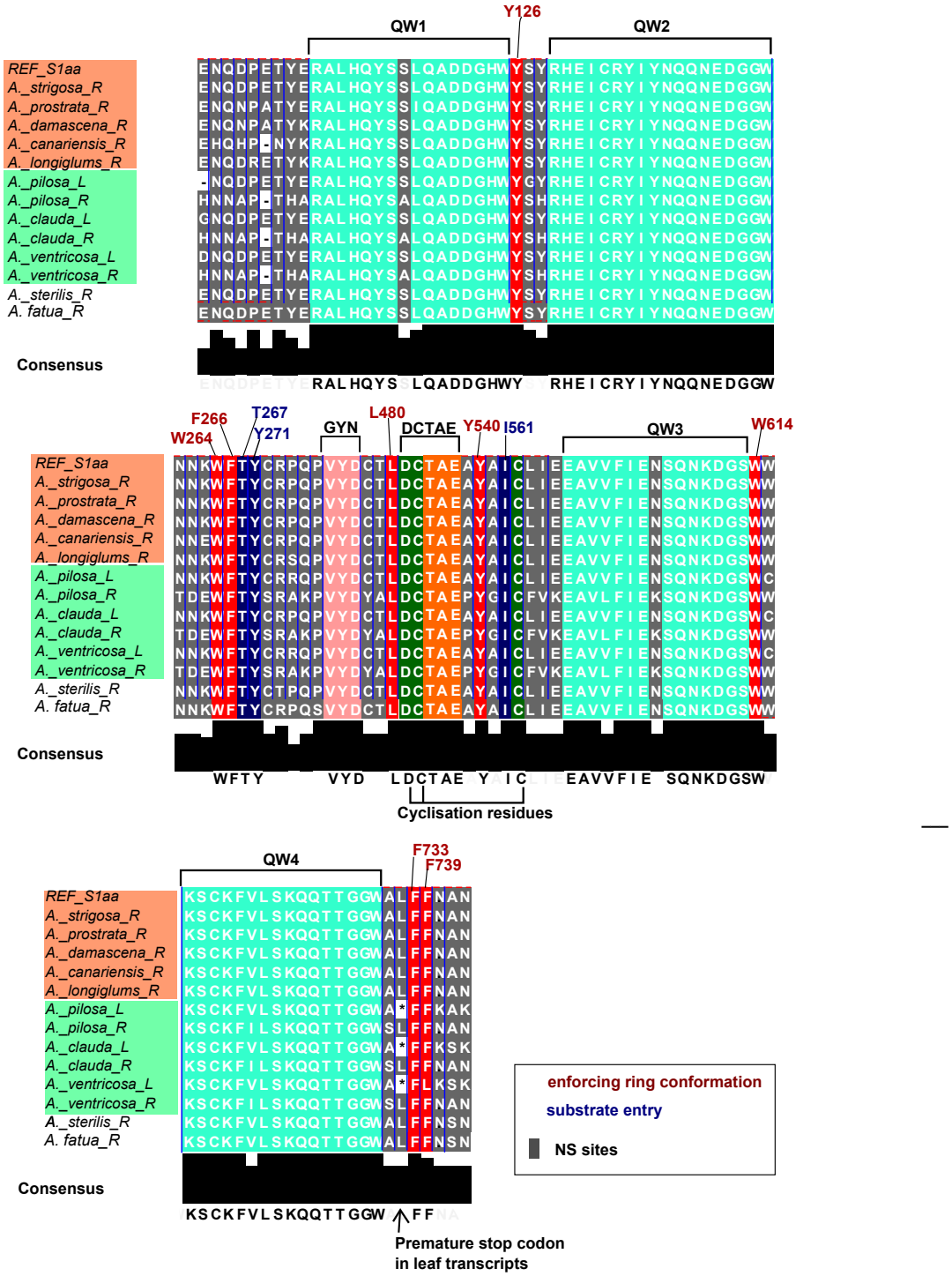


Figure 4.2: Minimal alignment of the predicted amino acid sequences for the oat *Sad1* homologues retrieved from RT-PCR. The QW, DCTAE, and GYN motifs are marked. Key active site residues for substrate entry (T267, Y271, and I561), those for enforcing ring conformation (Y126, W264, F266, L480, Y540, W614, F733 and F739), and those for cyclisation are indicated. Completely conserved residues are indicated underneath the alignment. NS sites are highlighted in grey and that where the premature stop-codons of the C genome oat leaf transcript sequnces located is arrowed. Leaf (L) and root (R) transcripts are labelled next to the species name. A and C genome species are highlighted in orange and green respectively. The complete alignment can be found in the Appendix (Alignment 4.1).

NS sites are located within the QW motifs (one in the QW1 and the other in the QW3 motif) and may have led to changes in product specificities in the C genome root expressed SAD1-like proteins (Figure 4.2). The leaf transcripts detected in the C genome oats contained an in-frame stop codon at 58 aa upstream of the conventional 3′ end (Figure 4.2), potentially producing a shortened version of the β-amyrin synthase protein. This will require verification by western blot analysis unsing antisera that are specific for SAD1.

Phylogenetic trees estimated for *Sad1* sequences using both maximum likelihood and Bayesian methods showed generally consistent tree topologies (Table 4.2) that the clade of C genome oat sequences separated from the A genome oats with an exception of *A. damascena* (Figure 4.3a to d and summary in Table 4.2), following the oat phylogeny. The root transcripts of *Sad1* homologues of *A. prostrata, A. canariensis, A. longiglumis* and *A. strigosa* grouped with their genomic sequences (with bootstrap supports 64, 100 and 71 respectively) in the codon trees (Figure 4.3b and d), even though these specific relationships are less clearly shown in the amino acid trees (Figure 4.3a and c), suggesting that the root transcripts were likely to originate from the genomic sequences of the corresponding species.

However, the sequences of *A. damascena*, which possesses the Ad genome, grouped in a markedly different way compared to the rest of the A genome oats. The root transcript of *A. damascena Sad1* homologue grouped with the A genome oat sequences as expected, while its genomic sequence grouped with the C genome oat genomic and root transcript sequences (bootstrap value = 100 in both Figure 4.3a and b) in all the *Sad1* trees (Figure 4.3a-d). This suggested that the root transcript of *A. damascena* is A genome-like and did not originate from the retrieved C genome-like genomic sequences.

The predicted amino acid sequences for the leaf transcripts detected in the C genome oats formed a distinct clade in the *Sad1* trees (bootstrap value = 87), while the C genome oat genomic sequences grouped with their root transcripts (bootstrap value = 100) (Figure 4.3a), suggesting that the root transcripts of the C genome oats originate from their genomic sequences of *Sad1* homologues while the leaf transcripts do not.

These results suggested that the root or leaf transcripts originated from paralogous *Sad1* genes rather than the retrieved genomic sequence. Alternatively, the differences between the genomic and the transcript sequences may be attributed to sequencing errors within the genomic sequences because of the difficulties of amplifications multiple PCR amplicons and direct sequencing of the template DNA (subsection 4.2.1).

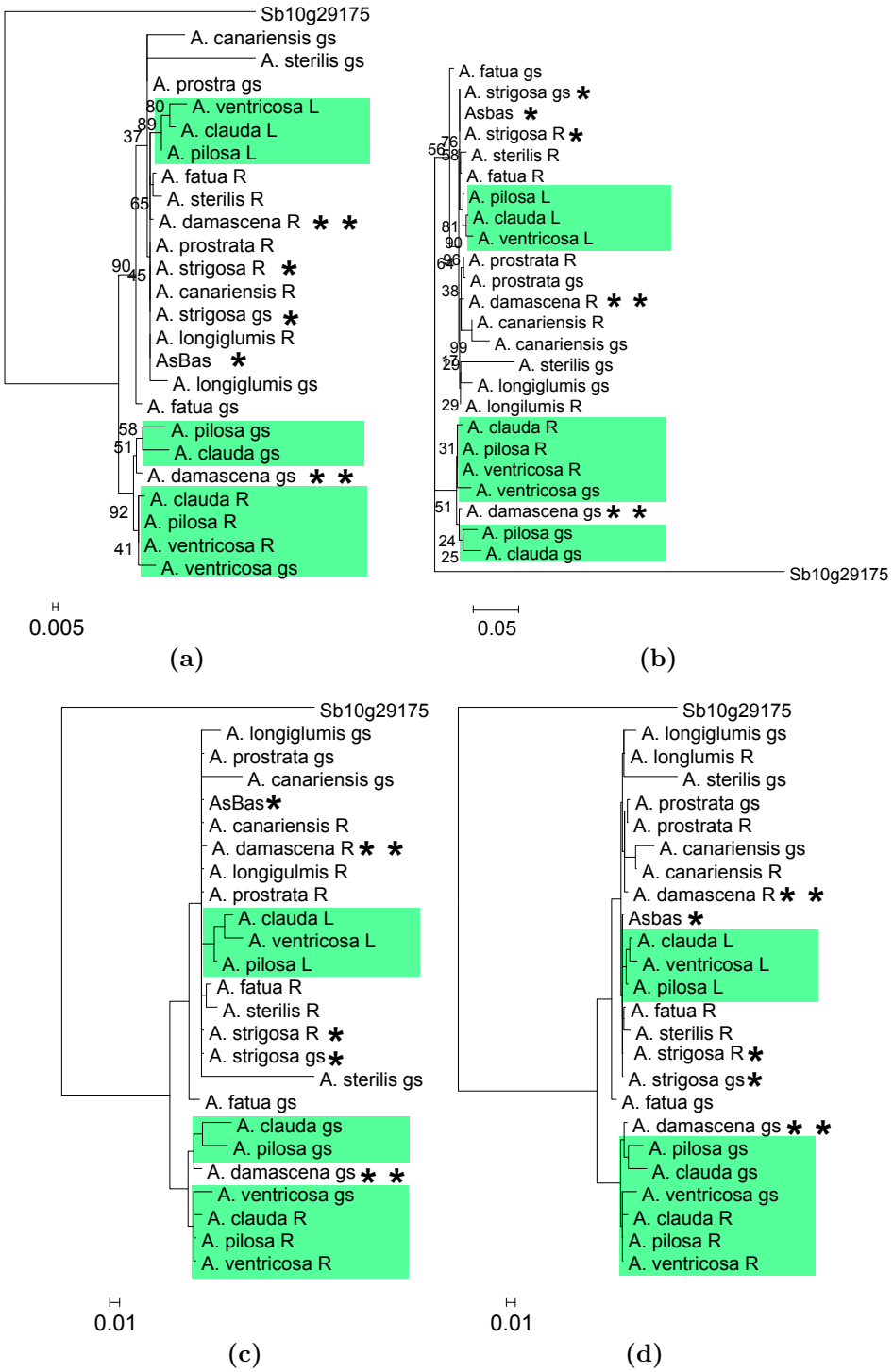


Figure 4.3: Phylogenetic trees of *Sad1* homologues from the analysed *Avena* species. The RAxML a) amino acid and b) codon alignment trees. The MrBayes c) amino acid and d) codon alignment trees. Bootstrap support for branches from 10,000 replicates is indicated for the RAxML trees. The posterior probabilities for branches from 75,000,000 MCMC samples in MrBayes trees are not shown. Tip label: reference sequence of *A. strigosa* SAD1 (Asbas), Genomic sequences (gs), Leaf (L) and root (R) transcript. *A. strigosa Sad1* sequences are marked with a single asterisk; the *A. damascena Sad1* sequeces are marked with two asterisks. The C genome oat sequences are highlighted in green. Sb10g29175 (*Sorghum bicolor Sad1* homologue) is used as the outgroup of the trees. The original tree files can be found in the Appendix (Tree 4.1 to 4.4).

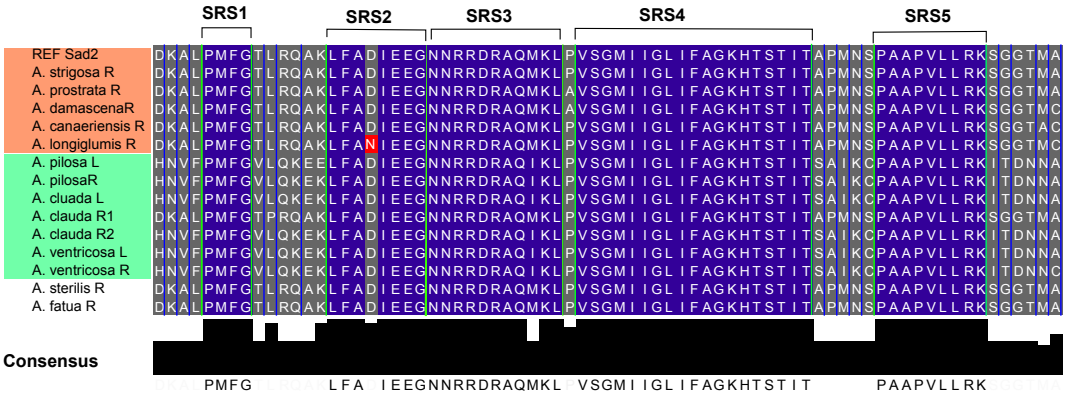


Figure 4.4: Minimal amino aicd alignment of the oat *Sad2* homologues retrieved from RT-PCR. The SRS1 to SRS5 motifs are marked. Completely conserved residues are indicated underneath the alignment. NS sites are highlighted in grey. The non-conserved residue (N instead of D) of *A. longiglumis* in the SRS2 motif is highlighted in red. Leaf (L) and root (R) transcripts are labelled next to the species name. A and C genome species are highlighted in orange and green respectively. The full-alignment can be found in the Appendix (Alignment 4.2).

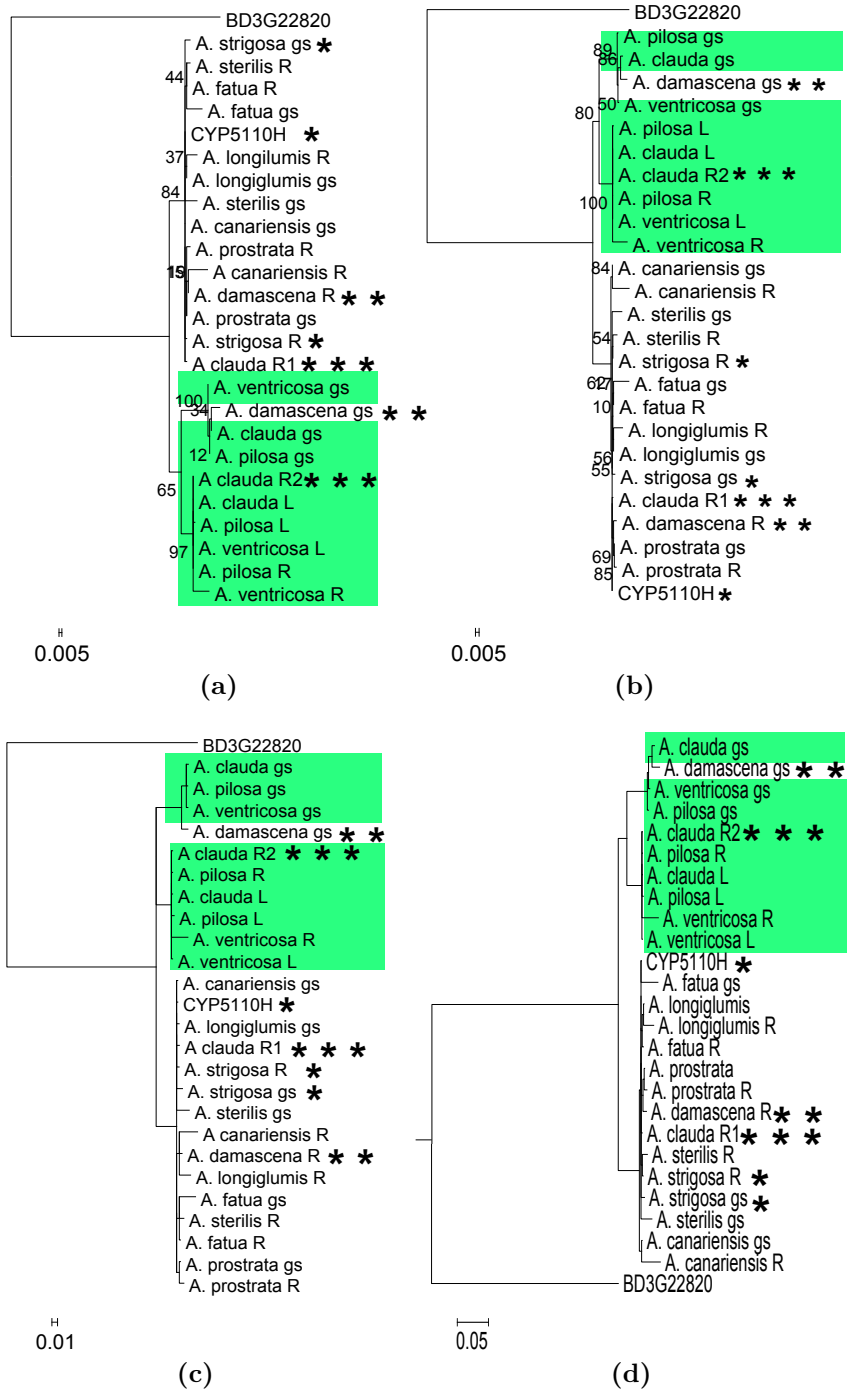


Figure 4.5: Phylogenetic trees of *Sad2* homologues from the analysed *Avena species*. RAxML tree from a) amino acid sequence alignment and b) codon sequence alignment. MrBayes tree from c) amino acid sequence alignment and d) codon sequence alignment. Bootstrap support for branches from 10,000 replicates is indicated for the RAxML trees. The posterior probabilities for branches from 75,000,000 MCMC samples in MrBayes trees are not shown. Tip label: Reference *Sad2* sequence (CYP51H10), *A. strigosa* sequences are marked with a single asterisk; the *A. damascena* sequences are marked with two asterisks; and the two root transcript of *A. clauda* are marked with three asterisks. Genomic sequences (gs), Leaf (L) and root (R) transcript sequences. C genome oat sequences are highlighted in green. BD3G22820 (*Brachypodium distachyon Sad2* homologue) is used as the tree outgroups. The original tree files can be found in the Appendix (Tree 4.5 to 4.8)

4.3.4. Analysis of oat *Sad2* homologues

The *A. strigosa Sad2* gene encode the enzyme AsCYP51H10, which modifies the triterpene scaffold during avenacin biosynthesis (Figure 1.2) (Geisler et al., 2013). Molecular modelling of AsCYP51H10 identified amino acid residues within the substrate recognition sites (SRS1-6) that are likely to be important for enzyme function and that together contribute to the dual reactions catalysed by this enzyme (C16 hydroxylation and C12, C13 epoxidation of β-amyrin) (Geisler et al., 2013). Amongst the *Sad2* homologue sequences, the SRS1 (PxFG, columns 115-118) , SRS2 (columns 204-211), SRS3(column 204-211), SRS4 (HT/sS, columns 291-293) and SRS5 (column 255-260) motifs were highly conserved (Figure 4.4) (Bellamine et al., 2004; Geisler et al., 2013; Lepesheva and Waterman, 2007) (Appendix Alignment 4.2). SRS6 was found to be located outside the highly conserved regions of the alignment and thus was not further analysed. In the alignment of *Sad2* homologue sequences, 24 NS sites were identified, 15 of which were differed between the A genome and the C genome *Sad2* homologue transcripts (Appendix Table 4.1 and Alignment 4.2). Of note, the *A. longiglumis Sad2* homologue possesses a NS difference within the SRS2 motif, which might alter the substrate recognition property of the gene product (Figure 4.4).

Phylogenetic estimations revealed that the transcripts of *Sad2* homologues of the A and C genome oats grouped into distinct clades (Figure 4.5a-d and summary in Table 4.2). Of the two transcripts recovered from the roots of *A. clauda*, one of them (*A. clauda* R1) grouped with the A genome oat sequences while the other (*A. clauda* R2) grouped with the C genome oat sequences (bootstrap value = 100 in both Figure 4.5a and b), suggesting that *A. clauda* may have two root-expressed *Sad2* homologues, of which *A. clauda* R1 is closely related to the A genome oat sequences. Unlike the *Sad1* trees, the leaf and root transcripts of the C genome oats were clustered (bootstrap value = 100 in both 4.5a and b), indicating that the leaf and root transcripts most likely originate from the same *Sad2* gene homologue, or alternatively from two independent but highly similar *Sad2* paralogues within the C genome oats. Interestingly, the amplified genomic sequence of the *A. damascena Sad2* homologue again clustered with C genome oat genomic sequences (Bootstrap value = 100 in Figure 4.5a and b) while its root-transcript sequence grouped with other diploid A genome oats, suggesting the root transcript does not originate from the genomic sequence that was amplified (Figure 4.5a-d).

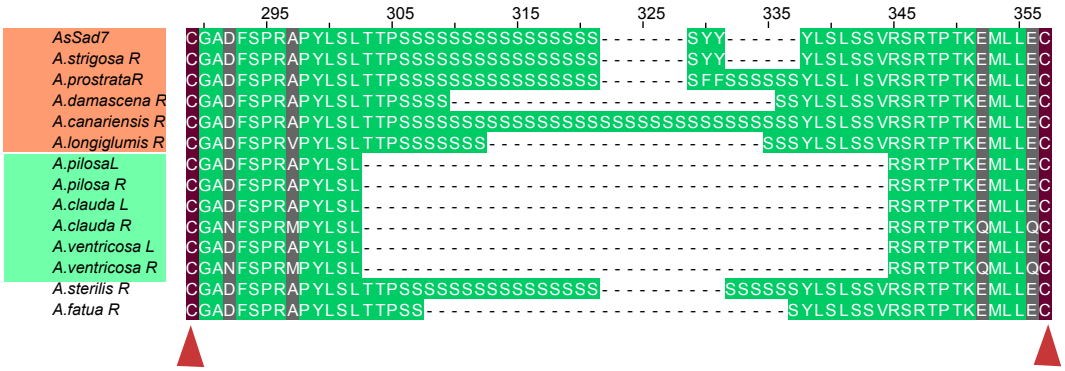


Figure 4.6: The minimal alignment of the predicted amino acid sequences of SAD7 from the transcripts of the examined *Avena* species in RT-PCR analysis showing the linker regions. The conserved cleavage sites (C288 and C343) are marked. NS sites within the linker region are highlighted in grey. Leaf (L) and root (R) transcripts are labelled with next to the species name. A and C genome species are highlighted in orange and green respectively. The full-length alignment can be found in the Appendix (Alignment 4.3).

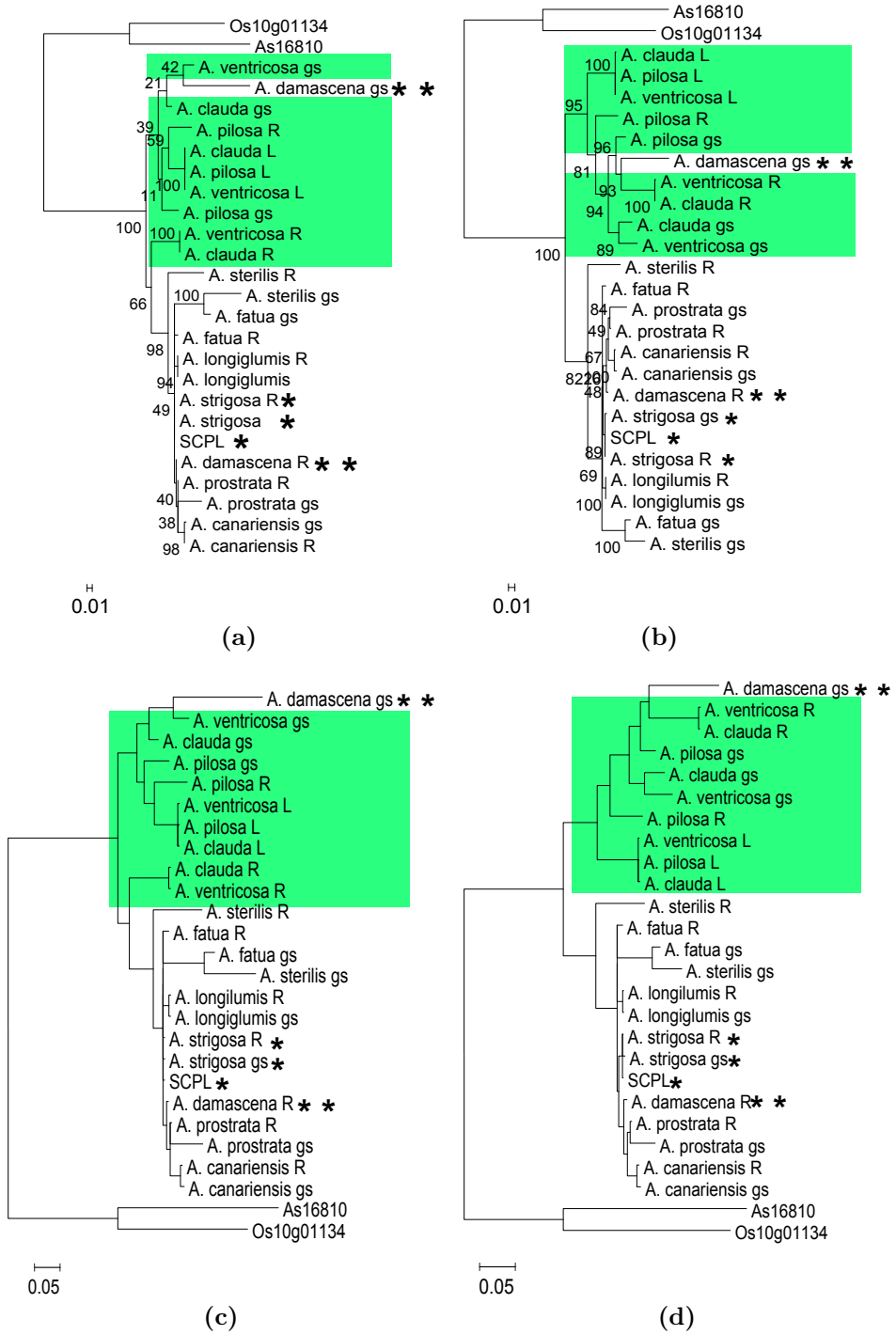


Figure 4.7: Phylogenetic trees of *Sad7* homologues from the analysed *Avena species*. RAxML tree from a) amino acid sequence alignment and b) codon sequence alignment. MrBayes tree from c) amino acid sequence alignment and d) codon sequence alignment. Bootstrap support for branches from 10,000 replicates is indicated for the RAxML trees. The posterior probabilities for branches from 75,000,000 MCMC samples in MrBayes trees are not shown. Tip label: Reference *Sad7* sequence (SCPL1), Genomic sequences (gs), Leaf (L) and root (R) transcript. *A. strigosa* sequences are marked with a single asterisk; the *A. damascena* sequeces are marked with two asterisks. C genome oat sequences are highlighted in green. Os10g01134 (*Oryza sativa Sad7* homologue) is used as the tree outgroup. The original tree files can be found in the Appendix (Tree 4.9 to 4.12)

4.3.5. Analysis of oat *Sad7* hommologues

Sad7 is a serine carboxypeptidase (SCPL) acyltransferase that is required for the addition of the acyl group (*N*-methyl anthranilate in the case of avenacins A-1 and B-1, and benzoate in the case of avenacins A-2 and B-2) (Figure 1.2). Sequence analysis of SCPLs revealed that SCPLs contain a highly conserved S-H-D catalytic triad, which serves as a charge relay system for acyl transfer (Stehle et al., 2006). Within the conserved regions in the *Sad7* sequence alignment (Appendix Alignment 4.3), two amino acid residues within this triad (S137, and D430) were found to be highly conserved. The third member of the triad (H483) lays outside of the conserved regions of the alignment and so was not examined. A total of 62 NS sites were detected amongst the *Sad7* sequences within the conserved regions of the alignment (Appendix Table 4.1).

A feature of SAD7 is the linker region that is postranslationally cleaved to give two subunits (Mugford et al., 2009). Although all *Sad7* homologue sequences contain the conserved splice sites (Mugford et al., 2009) of the linker regions, the length of the linkers is highly variable due to the differences in the number of serine residues within this region with *A. canariensis* SAD7 possessing up to 32 serine residues and the C genome *Sad7* sequences having none (Figure 4.6). The variation in the lengths of the linker region may not affect the SAD7 protein functions because it would eventually be removed from the mature proteins. However, there may be other functions of the linker of SAD7 in *A. strigosa* S75, such as regulation of translation rates (Mugford and Milkowski, 2012).

The *Sad7* phylogenetic trees showed consistent topologies (Figure 4.7a-d and summary in Table 4.2) that the A and C genome sequences formed separated clades generally. The leaf transcripts of the C genome oats formed a distinct clade (bootstrap value = 100 in Figure 4.7a and b) from the root transcripts and the genomic sequences of the C genome *Sad7* homologues, suggesting that the root transcripts are likely to originate from the genomic sequences but the leaf transcripts are not (Figure 4.7a-d). Of note, the genomic sequence of *A. damascena* was again clustered with the C genome oat sequences (bootstrap value = 93 in Figure 4.7b) while the root transcript sequence clustered with other diploid A genome oat sequences (bootstrap value = 68), suggesting the root transcript is likely to originate from the other *Sad7* homologue but not the retrieved genomic sequence (Figure 4.7a-d).

4.3.6. Sequence analysis of oat *Sad9* homologues

The *Sad9* sequences were highly diverse amongst the *Avena spp*, with 67 NS sites being detected in the sequence alignment (Appendix Table 4.1 and Alignment 4.4), of which 27 were accounted for non-synonymous differences between the A and C genome oat *Sad9* sequences.

4.3.7. Protein modelling of *Sad1*

Three dimensional model of root specific Sad1 constructed from the human lanosterol synthase (OSC) crystal structure

To assess the non-synonymous differences among the oat *Sad1* root transcripts, protein modelling was performed using I-TASSER 2.0.1 (Zhang, 2008) based on the human OSC crystal structures (PBD 1w6jA and 1w6k) (Thoma et al., 2004) (Appendix Table 4.2).
The protein models of the *Avena* SAD1 orthologues align to human OSC (1w6jk) with an average root-mean-square deviation (RMSD) of 0.6 $\AA$, indicating high structural similarities of these proteins (Appendix Table 4.2), differing mainly in the amino-terminal regions (1-80 aa) (Figure 4.8a, b and c). The H232 residue of human OSC was replaced by the aromatic residue F in all SAD1 proteins, which were speculated to facilitate E-ring cyclization (circled in Figure 4.9a1 and 2). In addition, the F444 in human OSC that stabilizes the orientation of the cyclisatioin intermediate after A-ring and B-ring formation (Thoma et al., 2004) is replaced by L in all SAD1 orthologues proteins (circled in Figure 4.9a1 and 2). Otherwise, all catalytic residues were conserved between protein models of human OSC 1w6k and SAD1s (Y98, W387, C456, D455, and C533) (Figure 4.9a and b).
The conformation of the C456-D455-C533 catalytic triad in the cyclisation domain are shown to be highly conserved amongst SAD1 proteins (Subsection 4.3.3; Figure 4.9). The non-synonymous differences identified previously in the multiple sequence alignment of *Sad1* sequences (Appendix Table 4.1) were mapped to the structural alignments (Appendix Alignment 4.5) of the SAD1 protein models. Eight of the 26 NS sites were located at the dynamic amino-terminal regions and are not expected to affect the protein conformation.
The membrane insertion regions were conserved in the multiple sequence alignment but not in the structural alignment, due to the highly variable conformations adopted by the loops encompassing the membrane insertion helix. Thus, the non-synonymous differences around the membrane insertion regions were not further investigated.

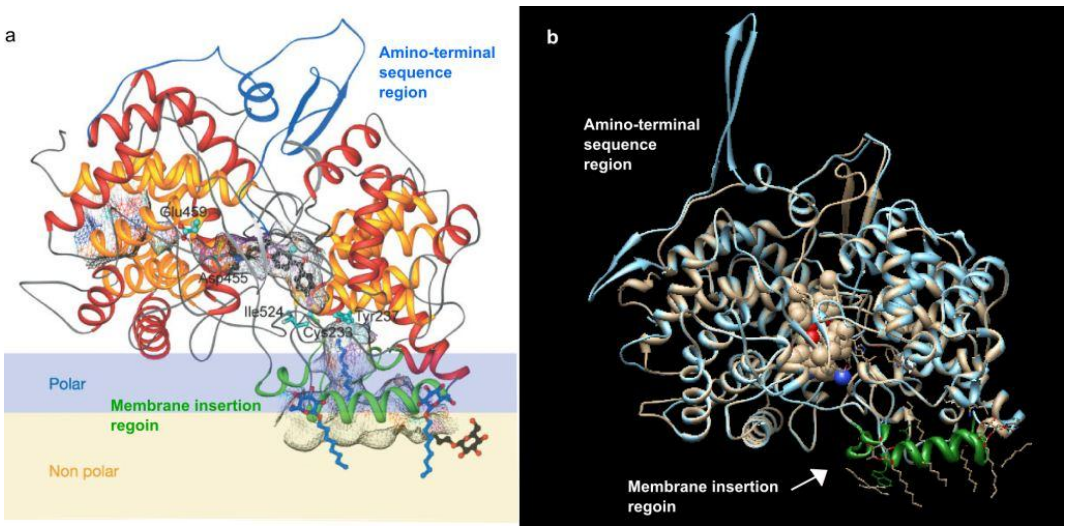


Figure 4.8: Protein models of OSC. a) The crystal structure of human OSC reproduced from Thoma et al. (2004) with different domains highlighted and the cellular orientation of the protein. b) Structural alignment of *A. strigosa* SAD1 (light blue) and human OSC (metallic). The amino-terminal region and membrane insertion domain (in green) are indicated. The structural alignment can be found in the Appendix (Alignment 4.5).

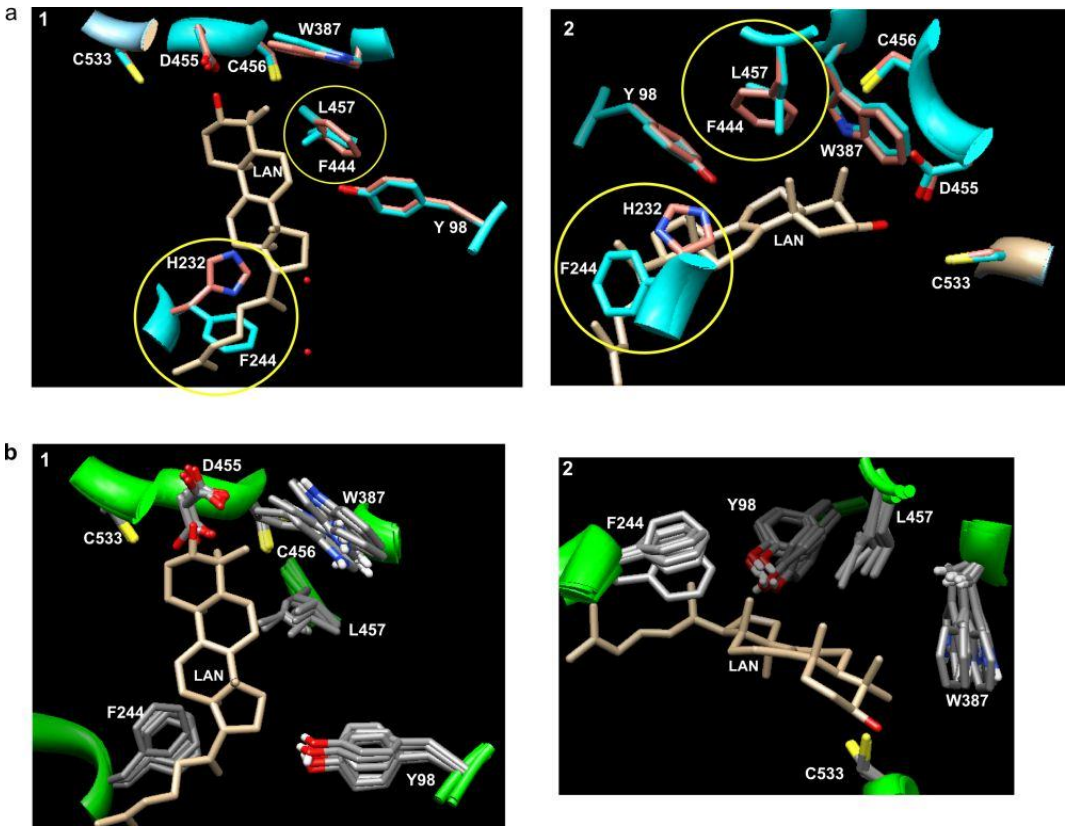


Figure 4.9: Syperposition of the active sites of SAD1 Structural alignment of a) the *A. strigosa* SAD1 (in blue) and human OSC (in red) (1w6jk) and b) oat SAD1 models focussing on the active site pocket. The key residues Y98, W387, D455, C456, C533 of human OSC are conserved in all SAD1 homologue proteins. However, the H232 and F444 in human OSC is replaced by F244 and L457 in SAD1 proteins (circled in a). 1) vertical view and 2) side view of the active sites. The structural alignment can be found in the Appendix (Alignment 4.5).

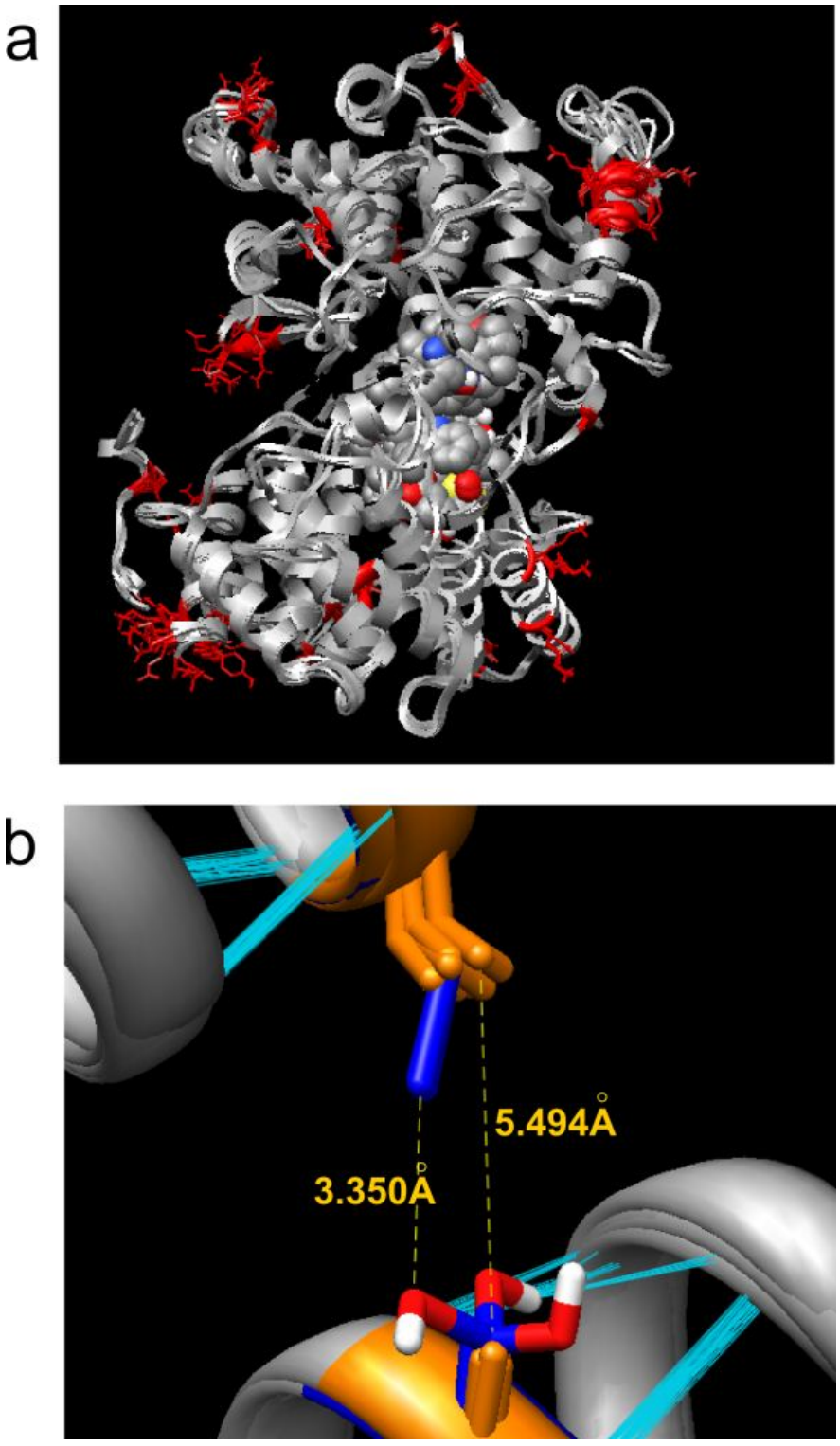


Figure 4.10: Structural alignment of SAD1 orthologues of all diploid oat species. a) the superposition of SAD1 orthologues. Conserved regions are in grey and the NS sites are highlighted in red. Active site residues are indicated with spheres. b) Structural alignment focussing on the likely co-evolved residues V/A pairs (in orange) present in the A genome and I/S pair (in blue) in C genome SAD1 models. Estimated hydrogen-bonds are highlighted in cyan. The estimated distance between the I/S pair (3.350$\AA$ indicated) is closer compared to that of the the V/A pair (5.494$\AA$).

The majority (32 out of 34) of non-synonymous differences among *Avena Sad1* amino acid sequences were located at the surface of the protein (Figure 4.10a). The non-synonymous differences that were buried inside the barrel were V630 and A669 of the *A. strigosa* S75 SAD1 amino acid sequence, which were conserved in all A genome oats but were replaced by I and S in the C genome species. Interestingly, these two sites face each other (Figure 4.10b). It could be speculated that the V-A pair may have coevolved to I-S for increased attraction between the two neighbouring α-helices.

Sites of non-synonymous differences were not free of constraints

While the non-synonymous differences among *Sad1* sequences (Appendix Alignment 4.5) were found to be mainly located on the protein surface (Figure 4.10a), the sites exhibiting these non-synonymous differences appeared to be under selective constraints of electrical charges. In the NS sites, small non-polar side chain residue A was replaced by S, G and P, while sulfur residue M was replace by C. Furthermore, charged residues were replaced by charged residues (N to K, N to H, and Q to K), regardless of the charges. It could be speculated that although sites at the protein surfaces are more tolerant to mutations (Wagner, 2008), there are still constraints on such changes in order to maintain the overall balance of surface charges and structural integrity.

4.4. Discussion

4.4.1. Sequence diversity of *Sad* gene homologues

Sequence analysis of *Sad1, 2, 7* and *9* revealed that the avenacin biosynthetic genes share high levels of sequence identity across the *Avena spp.* The consistent topologies shared between the re-estimated oat phylogeny (Figure 2.6a) and the gene trees of *Sad1, 2* and *7* (Figure 4.3a-d, 4.5a-d and 4.7a-d) suggested that the *Sad* genes are present in the ancestral *Avenastrum*, and have not evolved repeatedly in different oat species. The root transcripts of *Sad1, 2* and *7* detected in the RT-PCR analysis of the A genome oats (with the exception of *A. damascena*) all originate from the genomic sequences recovered from the corresponding species. The genomic and root transcript sequences of the C genome oats were distant from their leaf transcript sequences, which instead grouped with the A genome sequences, in the *Sad1* trees, suggesting that the leaf transcripts were likely to be orthologous to the A genome sequences while the root transcripts are A genome *Sad1* paralogues that have arisen from

duplication events preceding C genome oat radiation. In contrast, the leaf and root transcripts from C genome oats were clustered together in the *Sad2* trees, with very short branch lengths. Therefore, it may be that in the C genome oats, *Sad2* is expresses in both the roots and leaves, or alternatively that a paralogue highly similar to *Sad2* is expressed in the leaves of C genome oats. The leaf and the root transcripts of the C genome *Sad7* homologues grouped together in a distinct clade from the A genome sequences, suggesting that duplication of *Sad7* homologues in the C genome clade occurred after the A/C genome split. Therefore both C genome transcripts are co-orthologues of the A genome *Sad7*.

In the phylogenetic analysis of the *Sad* gene transcripts and genomic sequences, the *Sad1, 2* and *7* root transcripts of *A. damascena* grouped with other A genome oat sequences whereas the genome sequences recovered from this species for the three *Sad* genes repeatedly group with the C genome oat sequences. This suggested that there may also be multiple copies of *Sad1, 2* and *7* present in the *A. damascena* genome and that the root transcripts do not originate from the reconstructed genomic sequences.

Although species-specific non-synonymous differences were observed in homologues of *Sad1, 2, 7* and *9*, the active site residues and domains are well conserved, indicating that the products of the *Sad* genes of all oat species investigated are likely to be functional and that many sites are evolving under purifying selection.

In order to evaluate the significance of these non-synonymous differences on protein functionality, three-dimensional structural alignment of SAD1 protein models were built with the non-synonymous differences were mapped on to it. Most of the non-synonymous differences were found to be located at the protein surface or within loops between helices, which are robust to changes (Masel and Trotter, 2010). Nonetheless, the non-synonymous changes seemed to be non-random so as to maintain the overall charges on the protein surface. This suggests the non-synonymous differences among oat *Sad1* coding sequences may have occurred at 'neutral' sites within a protein evolving under purifying selection, so as to maintain the overall protein conformation. However, protein modelling was not able to further assess precisely the impacts on protein stability or protein-protein interactions brought about by these non-synonymous differences.

In the future, protein modelling will be extended to *Sad2, 7,* and *9*, which contain more inter-specific non-synonymous differences than *Sad1* and relatively flexible protein folds once X-ray protein structures in eukaryotes are available.

4.4.2. *Sad* gene evolutionary history relative to oat phylogeny

The experiments reported here collectively show that all *Avena* species investigated possess a set of *Sad1, 2, 7* and *9* genes responsible for avenacin biosynthesis, as an inherited feature from the ancestral *Avenastrum*. The leaf transcripts detected in the C genome oats may originate from *Sad* gene copies that have arisen in the ancestral C genome oat after the A and C genome split. Alternatively, the leaf expressed *Sad1, 2* and *7* homologues of the A genome oats may have been lost after the A and C genome split.

The data presented in this chapter again show that the avenacin biosynthetic genes are highly conserved within the A and C genome oat species respectively, but not between the two different genome types, suggesting that these gene sets have undergone independent evolution in different *Avena* species following the divergence from the ancestral *Avenastrum*.

Chapter 5 - Phylogenetic studies of *Sad* genes

5.1.Introduction

In this chapter, the sequences of the five *Sad* genes are analysed using a phylogenomic approach similar to those of Inagaki et al. (2011) and Xue et al. (2012). This approach enables founder events of avenacin biosynthesis within the *Poaceae* to be identified.

5.1.1.Development of a phylogenetic pipeline to investigate the evolution of triterpene biosynthetic genes in monocots

Evolutionary analyses of the oxidosqualene cyclase (OSC) and cytochrome P450 51 (CYP51) triterpene biosynthetic genes were first performed in the sequenced rice genome (Inagaki et al., 2011). The workflow used was subsequently adapted for a wider investigation of OSC genes in plants (Xue et al., 2012) (Figure 5.1). Genome mining of *Sad1* and *Sad2* homologues of *A. strigosa* in rice via BLAST analysis (Gish, 1994) uncovered 12 *OSC* sequences and 12 *CYP51* sequences (Inagaki et al., 2011). Phylogenetic analysis revealed that most of the rice *OSC* genes had diverged via duplication from the ancient monocot cycloartenol synthase (CAS) following the dicot/monocot spit (Inagaki et al., 2011). The only exception was the *O. sativa OSC12*, which was likely to have derived from an ancient duplication of an ancestral *CAS* gene before the dicot/monocot divergence (Inagaki et al., 2011). Branch-site selection tests (Yang, 2007) were carried out on the phylogenetic trees of *OSC* and *CYP51*. The results of these tests showed that purifying selection had dominated the amino acid evolution of these two classes of genes. Furthermore, high levels of synonymous changes, potentially driven by an independent selective process had occurred (Inagaki et al., 2011). Examination of the chromosomal locations of the rice *OSC* and *CYP51* genes in the *Oryza sativa* genome sequence showed that *OsOSC1* and *OsCYP51H5* genes are in proximity (1.4 MB apart on rice chromosome 2) but

no evidence of triterpene gene clustering was uncovered (Inagaki et al., 2011).
Using a similar computational approach, Xue and co-workers subsequently performed a broader genomic search, identifying 96 *OSCs* in higher plants. They also showed via phylogenetic analysis that these *OSCs* had arisen from an ancient CAS gene (Xue et al., 2012). Through aligning the *OSC* phylogenetic tree to major duplication and speciation events known to have occurred in the evolution of land plants, the timing of the key duplication event giving rise to the ancient cycloartenol synthase (CAS), that had led to the birth of monocot triterpene synthases, was inferred to be approximately 140 mya (Xue et al., 2012). This ancient CAS gene underwent gene family expansion in monocot lineages via a tandem duplication event prior to the ρ whole genome duplication (WGD) event (Xue et al., 2012). Studies of branch specific *dN/dS* ratios indicated relaxed selection on one of the two sister branches after a duplication event, confirming that *OSC* evolution is through duplication-neo-functionalisation (Xue et al., 2012).
It has been shown previously that *Sad1* and *2* have arisen by duplication and divergence of genes involved in sterol biosynthesis (Haralampidis et al., 2001; Qi et al., 2006). Both phylogenetic studies discussed above further confirmed that the *OSCs* involved in secondary metabolism largely evolved independently in dicots and monocots through duplication-neo-functionalisation from *OSC* genes involved in primary metabolism (Inagaki et al., 2011; Xue et al., 2012).
In this chapter, the previous phylogenetic pipeline (Figure 5.1) (Inagaki et al., 2011) was refined in order to increase the sensitivity of homologue searches and to enhance the accuracy of phylogenetic analysis of the monocot triterpene biosynthetic genes. The updated pipeline was first developed for phylogenetic analysis of the gene family that *Sad7* belongs to and was subsequently used to analyse gene family members of *Sad9*, *Sad10*, *Sad1* and *Sad2*.

5.2.Materials and methods

5.2.1.Retrieval of *Sad* gene homologues

BLAST (Gish, 1994) analysis was carried out, using the protein sequence of the relevent *Sad* gene in *A. strigosa* S75 as the query, against the protein sequence databases of *Oryza sativa, Zea mays, Sorghum bicolor,* and *Brachypodium distachyon* and the wheat cDNA database from TriFLBD (Table 5.1). Genomic sequences of the closest BLASTp hits (Gish, 1994) were downloaded from the databases. The coding sequences and translated amino acid sequences of the various genes were annotated manually using the Wise2 (EBI sever)(Birney

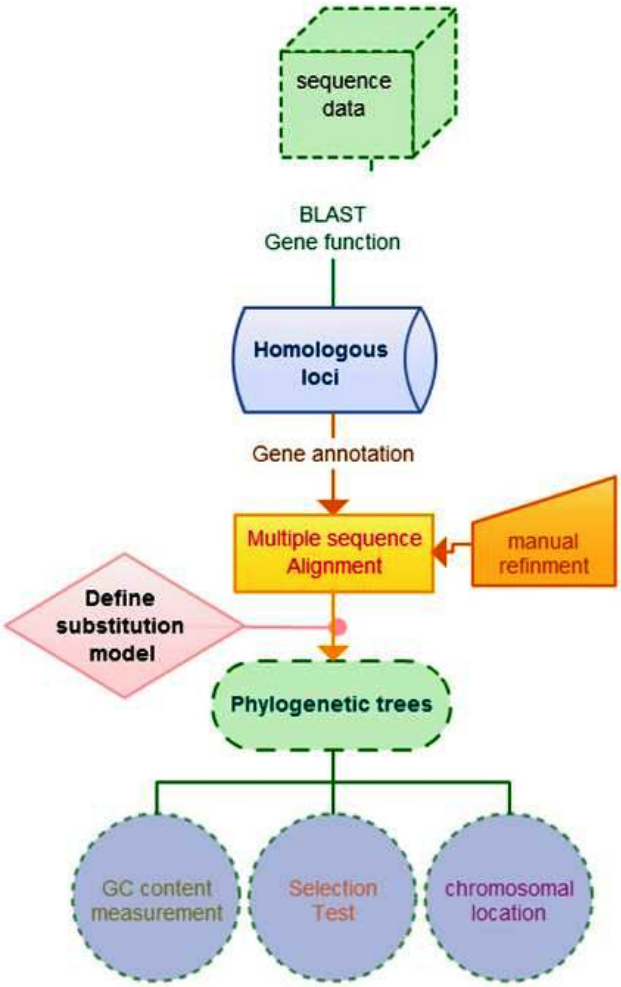


Figure 5.1: Summary of the phylogenetic workflow. Similar sequences were first recovered via BLAST analysis or gene ontology search. The coding sequences of the corresponding loci were checked or annotated manually followed by multiple sequence alignment. Phylogenetic tree estimation was carried out using RAxML 7.0.3 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) with the manually refined aligments and substitution models selected by ProtTest (Abascal et al., 2005) and FindModel (Posada and Crandall, 2001). Selection tests using PAML 4.5 (Yang, 2007), syntenic mapping and gene GC content measurement were carried out and these results were analysed within the framework of the phylogenetic trees.

et al., 2004) and Geneseqer software (Plant GDB server)(Brendel et al., 2004), with the relevant *A. strigosa Sad* gene coding sequence as a template. The resulting manually annotated sequences were then aligned using MUSCLE 3.6 (Edgar, 2004). Gaps were retained and the alignment was converted to Stockholm format and input to HMMER 3.0 (Eddy, 2009).

5.2.2.HMMER search and initial BioNJ tree construction

A Hidden Marhov Model (HMM) profile was constructed from the gapped alignment of the closest hits of the relevant *Sad* gene. Using the HMM profile, an HMM search of the peptide databases listed in Table 5.1 was performed locally using the NBI Computational Biology Cluster. Prior to searching, the cDNA databases were translated to peptide sequence using the six-frame translation function of WU-BLAST (Gish, 1994). Hits retrieved from HMMsearches were mapped to previously published gene family trees. E-value thresholds of gene family clades selected for subsequent analysis were determined in one or more species (e.g. the E-value 1 x e^{-50} distinguished rice SCPL acyltransferases (Clade1A) from other clades in the *Sad7* analysis). Translated peptide sequences of the filtered HMMsearch hits were downloaded directly from the databases (Table 5.1) and were aligned using MUSCLE 3.6 (Edgar, 2004). Truncated sequences in the amino acid alignment were removed, followed by removal of highly variable regions (columns that contain more than five different amino acid residues). Finally, two different alignments were made. In the minimal alignment, no gapped columns were removed. In contrast, gapped columns were removed from the minimal alignment to create the stripped_column alignment. Neighbor joining trees were constructed from both alignments using the BioNJ method in SplitsTree4 (Huson and Bryant, 2006). Tree topologies of both trees were compared for consistency and the tree obtained from the minimal alignment was used to guide gene sequence annotation (i.e. choice of sequence template). Sequences that grouped within the chosen clade in the minimal alignment tree were selected for gene annotation and further phylogenetic analysis.

5.2.3.Gene annotation and gene structure determination

Sequences of *Sad* gene homologues in *A. thaliana* and *Physcomitrella patens* were downloaded directly from TAIR 9 and Phytozome v8.0 respectively; for alternatively spliced genes only the longest spliced form was retained. The genomic sequences of the selected monocot genes were manually annotated in

Wise2 (Birney et al., 2004) and GeneSeqer (PlantGDB) (Brendel et al., 2004), using the annotated peptide sequence of the most closely related gene in the same species as a gene model template. Alternatively, if a (clade-specific) gene family member in the same species was not found, the annotated peptide sequence of the most closely related gene in the minimal alignment tree was used instead. The annotated coding sequences of *O. sativa* and *Z. mays Sad* genes were compared to their corresponding full-length cDNAs (fl-cDNAs) using NCBI BLASTn (Gish, 1994), with the predicted coding sequence as a query in each case. Any discovered mismatches between the predicted annotated sequence and the fl-cDNA sequence were corrected according to the fl-cDNA.

The gene co-ordinates, gene structure details, reference cDNA accession numbers, and descriptions of gene function (if available) for each *Sad* gene homologue included in the phylogenetic analysis are listed in Appendix Table 5.1. Exon numbers and the ten amino acids spanning each exon junction (5 aa at 5′end and 5 aa at 3′end) of each annotated gene are also listed and the predicted gene structures are examined for: (1) conservation of gene structure, and (2) conservation of amino acids that span exon junctions, for validation of correct gene annotation. The translated amino acid and coding sequences of each gene are listed in Appendix Table 5.2 together with a description of any frame-shift mutations, amino acid deletions, truncations or mismatch(es) compared to the reference full-length cDNA.

5.2.4.Multiple sequence alignment

For each *Sad* gene, all manually annotated and downloaded peptide sequences were aligned using MUSCLE 3.6 (Edgar, 2004) to generate a amino acid alignment. The analogous coding sequences were aligned to the amino acid alignment using the PAL2NAL web sever (Suyama et al., 2006) in order to generate a coding sequence alignment. Both amino acid and coding sequence alignments were then manually refined in BioEdit (Hall, 1999) to remove highly variable regions (>5 different amino acids in a column) and all gapped columns.

Codon alignments were further processed in R2.15.0 (R Core Team, 2012) with the seqinr package (Charif and Lobry, 2007) to remove columns corresponding to third codon positions. This generated a 'third_base_strip' alignment, useful for phylogenetic analysis when inconsistencies between phylogenetic trees estimated from amino acid and coding sequence alignments were observed.

Species	Database version	Source of database	Database type	Species abbreviation
Arabidopsis thaliana	Arabidopsis_thaliana TAIR9.pep.all	Plantensembl (Kersey et al., 2012)	annotated genome	AT
Brachypodium distachyon	Brachy_ pep1.0. pep.all.	Plantensembl (Kersey et al., 2012)	annotated genome	BD, BRADI
Oryza sativa	O. sativa MSU6 pep all	Plantensembl (Kersey et al., 2012)	annotated genome,	LOC_Os, Os
Sorghum bicolor	S. bicolor.Sbi1.pep.all	Plantensembl (Kersey et al., 2012)	annotated genome	Sb
Physcomitrella patens	Phypa1.1.pep.all	Plantensembl (Kersey et al., 2012)	annotated genome	Pp
Zea mays	Z. mays AGPv2.pep.all	Plantensembl (Kersey et al., 2012)	annotated genome	GRMZM, ZM
Setaria italica	Sitalica_164_peptide	Phytozome v.8.0 (Goodstein et al., 2012)	annotated genome	Si
Avena strigosa root specific cDNA database	Oat_454_over1000bp contigs	The Genome Analysis Centre (TGAC)	fl-cDNA	As
Triticum aestivum (wheat)	RIKEN Triticum aestivum fl-cDNA	TriFLBD (Mochida et al., 2009)	fl-cDNA	Tri
Hordeum vulgare	BARLEY DB Hordeum vulgare fl-cDNA	TriFLBD (Mochida et al., 2009)	fl-cDNA	AK

Table 5.1: The genomic databases used in the Sad gene homologue searches. The species abbreviations of sequences retrieved from the different databases are listed.

5.2.5.Phylogenetic tree estimation

Appropriate substitution models for phylogenetic analysis were predicted for both alignments of each *Sad* gene using ProtTest (Abascal et al., 2005) (for peptide alignments) and FindModel (Posada and Crandall, 2001) (for coding sequence alignments). The criterion for model selection was the minimal Akaike information criterion (AIC) value. Phylogenetic trees were estimated from the alignments, with the predicted models, using RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001). In each RAxML analysis, the best-likelihood (BLK) tree was found from 50 or 100 maximum likelihood tree constructions. Bootstrap values were subsequently obtained from 10,000 bootstrap trees that were mapped back to the BLK tree. In the MrBayes analyses, phylogenetic trees were constructed from 100,000,000 MCMC simulations, with a burnin factor of 0.25 and sampled every 1,000 bootstraps until the posterior probability divergence of the sampled tree topologies <0.05. The 50 majority consensus tree was then summarized from the 75,000,000 retained MCMC samples. For each *Sad* gene, the consistency of the tree topology across the four phylogenetic trees estimated in RAxML 7.0.4 (Stamatakis, 2006) and MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001) was assessed. When the bootstraps value of a node of interest in a phylogenetic tree was low, or an inconsistency between trees generated from the amino acid and coding sequence alignments was observed, other sets of phylogenetic trees were generated (e.g. third_base_strip trees or trees estimated for close homologues only).

5.2.6.Molecular evolution analysis

Selection tests were carried out on the coding sequence alignments of each *Sad* gene. Branch-site tests were carried out in PAML 4.5 (Yang, 2007) on selected branches of the phylogenetic trees constructed from the amino acid alignments in RAxML 7.0.4 (Stamatakis, 2006). Branches evolving under positive selection were identified by performing likelihood ratio tests (LRTs) between Model M1A (neutrally selected foreground branch) and M2A (positively selected foreground branch) (Yang and dos Reis, 2011) at a 5% significance level, with a Bonferroni correction when multiple branches had been tested. Pairwise *dN/dS* values of coding sequences were calculated using CODEML with runmode = -2. The GC and GC_3 contents of each gene sequence in the coding sequence alignments were measured using custom R scripts that called the seqinr package (Charif and Lobry, 2007) in R2.15.0 (R Core Team, 2012).

5.3.Results of the *Sad7* phylogenetic analysis

Sad7 is a member of a multigene family consisting of three main functionally distinct groups: the serine carboxy peptidases and two groups of serine carboxy peptidase-like acyltransferases (SCPs, the SCPL Clade1A and Clade1B) (Mugford et al., 2009). Sequence analysis and catalyic activity assays have confirmed that SCPs and SCPL acyltransferases employ the same catalytic triad for enzyme activities (Stehle et al., 2006). The SCPLs were recruited from peptide bond hydrolysis to acyl transfer after the divergence of higher plants from mosses, and the monocot SCPLs have evolved separately from the dicot SCPLs (Mugford et al., 2009) (Figure 5.3).

5.3.1.Sequence retrieval and annotation of *Sad7* homologues

The BLASTp (Gish, 1994) search identified LOC_Os10g01134, BRADI3G21550, Sb01g027540, wheat contig Tri950, and GRMZM2G179528 as the closest hits to the *A. strigosa Sad7* translated peptide sequence in *O.sativa, B. distachyon, S. bicolor*, wheat and *Z. mays* respectively. Following annotation of these five loci, a *Sad7*_closest_hits HMM profile was built and used to identify a total of 254 hits from the surveyed species (Table 5.1). The two BioNJ trees constructed from the minimal_alignment and strip_column alignment (Trees 5.1 and 5.2 in the Appendix) exhibited consistent tree topologies. In general, the BioNJ trees consisted of three main clades (SCP, 120 sequences; SCPL Clade1A, 115 sequences; SCPL Clade1B, 19 sequences), similar to *O. sativa* sequences in the published SCPL tree (Figure 5.3). *P. patens* peptides all grouped within the SCP clade, suggesting that SCPL genes are restricted to higher plants. The SCPL Clade1A (Figure 5.2) was found to contained a dicot subclade (24 *A. thaliana* sequences) and a monocot subclade (86 sequences, including 10 Sad7 orthologues), suggesting that the expansion of the SCPL family occurred after the monocot-dicot split. An intermediate SCPL group of 5 monocot sequences was also observed, basal to Clade1A but showing a closer relationship to Clade1A than to Clade1B (Figure 5.2).

The 91 monocot genes in SCPL Clade1A were manually annotated while the 24 dicot sequences were downloaded directly from TAIR10 (TAIR, http://arabidopsis.org) (Lamesch et al., 2012). Gene annotations revealed that the exon that encodes the linker region of the SCPL, and its two flanking exons, are highly variable regions. Comparison of fl-cDNAs and predicted coding sequences in *O. sativa* and *Z. mays* genes indicated that sequence prediction

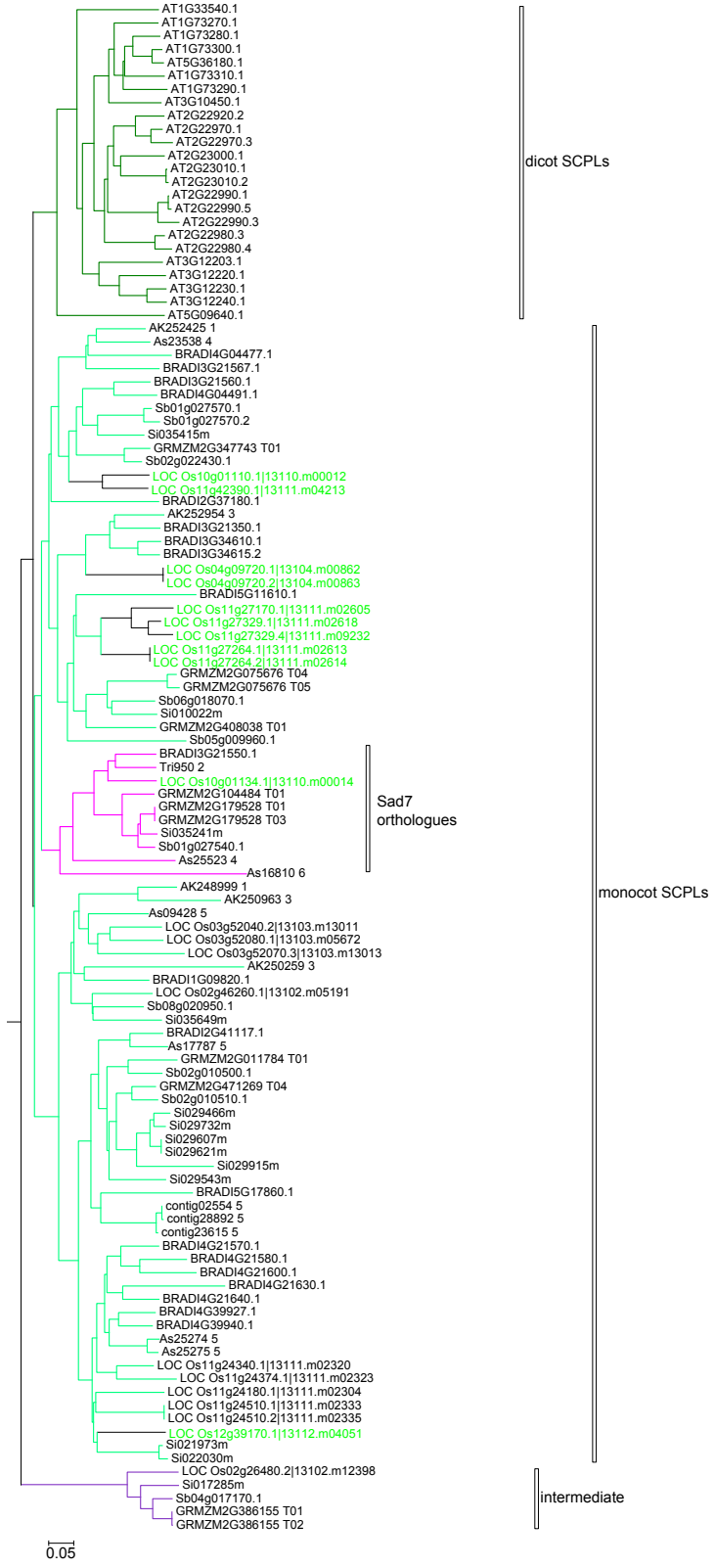


Figure 5.2: The BIONJ subtree constructed from the minimal amino acid alignment of SCPL Clade1A sequences. The basal group of five monocot genes (intermediate group) are defined to be either SCPL Clade1A or Clade 1B with reference to the published SCPL tree (Mugford et al., 2009). The genes that are also present in the reference SCPL tree are labelled in green. The gene IDs follow the labelling scheme of Table 5.1.

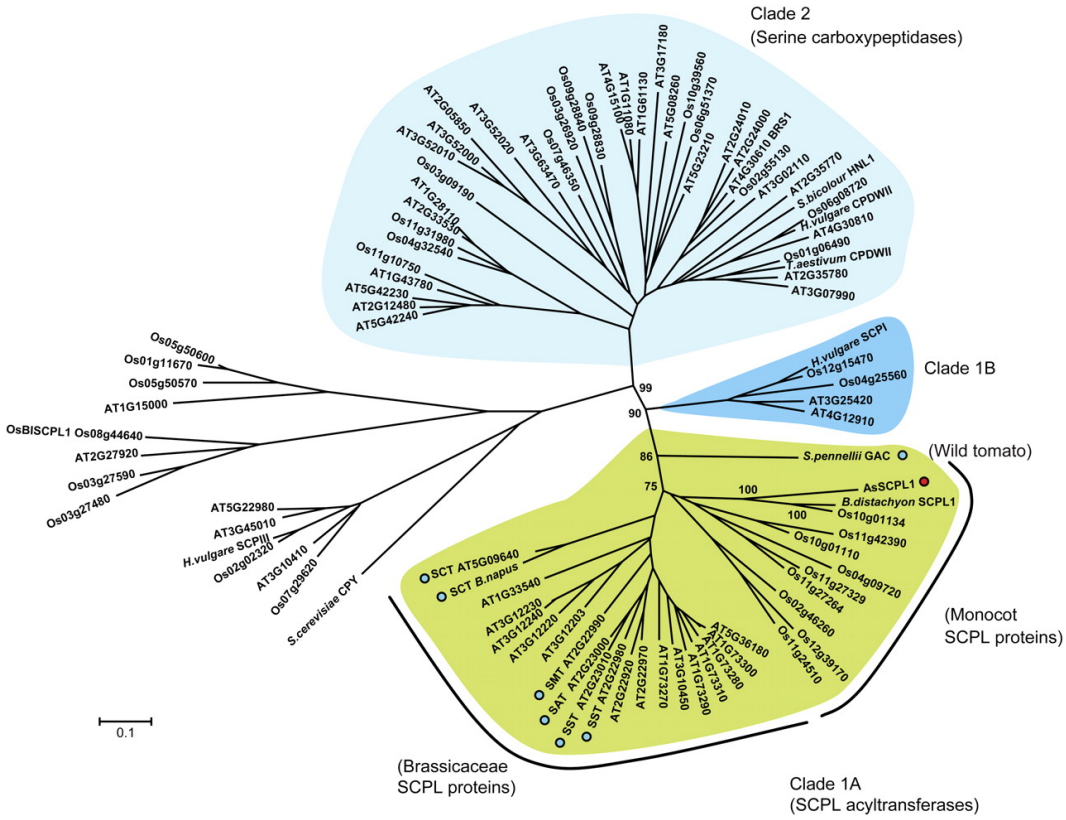


Figure 5.3: The neighbor-joining tree of plant SCPLs and SCP protein sequences reproduced from Mugford et al. (2009).

within this region was often inaccurate. Consequently, fl-cDNAs were used to guide coding and translated peptide sequence annotation within this highly variable region.

The intron-exon structure of each annotated gene was recorded in Appendix Table 5.1 and the annotated coding and translated peptide sequences were listed in Appendix Table 5.2, along with a description of annotation details. The exon analysis indicated that most genes in monocot Clade1A possess either 12 or 14 exons (Figure 5.4). However, the exon number of the *Sad7* orthologues was remarkably reduced, with BRADI3G21550 possessing only a single exon, LOC_Os10g01134 containing three exons, and Sb01g027540, Si035241, GRMZM2G104484, and GRMZM2G179528 all exhibiting a consistent structure of six exons (Figure 5.4 and Appendix Table 5.1). Notably, genes with reduced exon number, except the truncated genes, tended to have lost multiple adjacent introns. Nonetheless, the merged exons retained their conserved exon junctions.

Summary of SPCL Clade1A phylogenetic trees						
Phylogenetic software	Alignment type	No. sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	86	282 aa	-17956.96	1.30	JTT + I + Γ
RAxML	codon	86	849 bp	-36390.53	0.89	GTR + Γ
MrBayes	amino acid	86	282 aa	-18081.32	0.12	JTT + I + Γ
MrBayes	codon	86	849 bp	-36497.88	0.00	GTR + Γ

Table 5.2: Summary of SCPL Clade 1A phylogenetic analyses

5.3.2.Multiple sequence alignment and phylogenetic tree estimation of SCPL Clade 1A sequences

During the multiple sequence alignment step, the highly variable N-terminal signal sequence and linker regions were removed. LOC_Os12g27170, Sb05g09969, BRADI4G21580, and AT1G33540 were also removed from the alignment due to their short length. Finally, columns containing gaps were removed, resulting in the final alignments. The selected evolutionary models for these alignments, or the next more complex model if not present in the software, are shown in Table 5.2. The four phylogenetic analyses gave highly similar topologies, differing mainly in the grouping of BRADI2G37180 (Figures 5.5 and 5.6). The *A. thaliana* and monocot SCPLs were clearly separated into distinct clades, with 100% bootstrap support.

A group of sequences paralogous to those within the *Sad7* group was discovered, including sequences from *S. bicolor, S. italica, B. distachyon* and *O. sativa*. In order to obtain a clearer understanding of the topologies of these two groups, the 23 sequences were subjected to a further phylogenetic analysis, with AT5G09640 as an outgroup (Figure 5.7). The two tree topologies were highly similar, but with BRADI3G31780 differently grouped between the MrBayes and RAxML trees. The *Sad7* orthologous clade and the tandem paralogous clade both follow the pattern of the grass phylogeny, suggesting that an ancient tandem duplication event gave rise to the *Sad7* orthologues before the *Poaceae* radiation, and that both copies have been preserved throughout the evolution of the grasses. The analysis also identified a very closely related *Sad7* paralogue in oat, As16810, from the oat root cDNA library.

5.3.3.Molecular evolution analysis of *Sad7* homologues

Branch-site tests were performed on ten branches of interest in the *Sad7* orthologue/tandem paralogue tree (Figure 5.8). Each branch was tested for M1A (neutrally selected foreground branch) or M2A (positively selected foreground branch). Branch-site tests with PAML4.5 (Yang, 2007) indicated

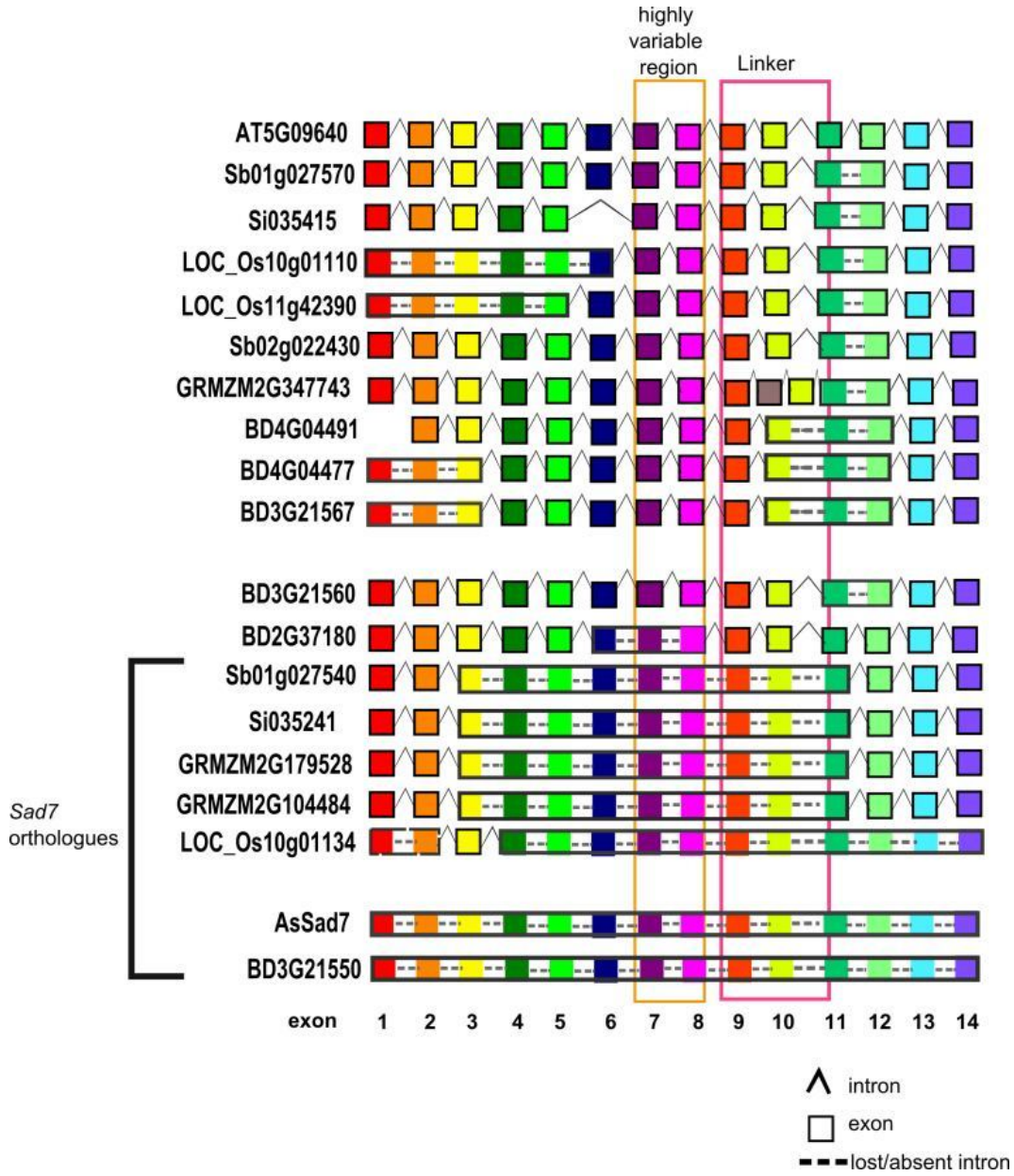


Figure 5.4: SCPL Clade1A gene structures. The coding sequences corresponding to exons 1 to 14 of the 14-exon sequence AT5G09640 are colour-coded. The *Sad7* orthologues have experienced intron reduction.

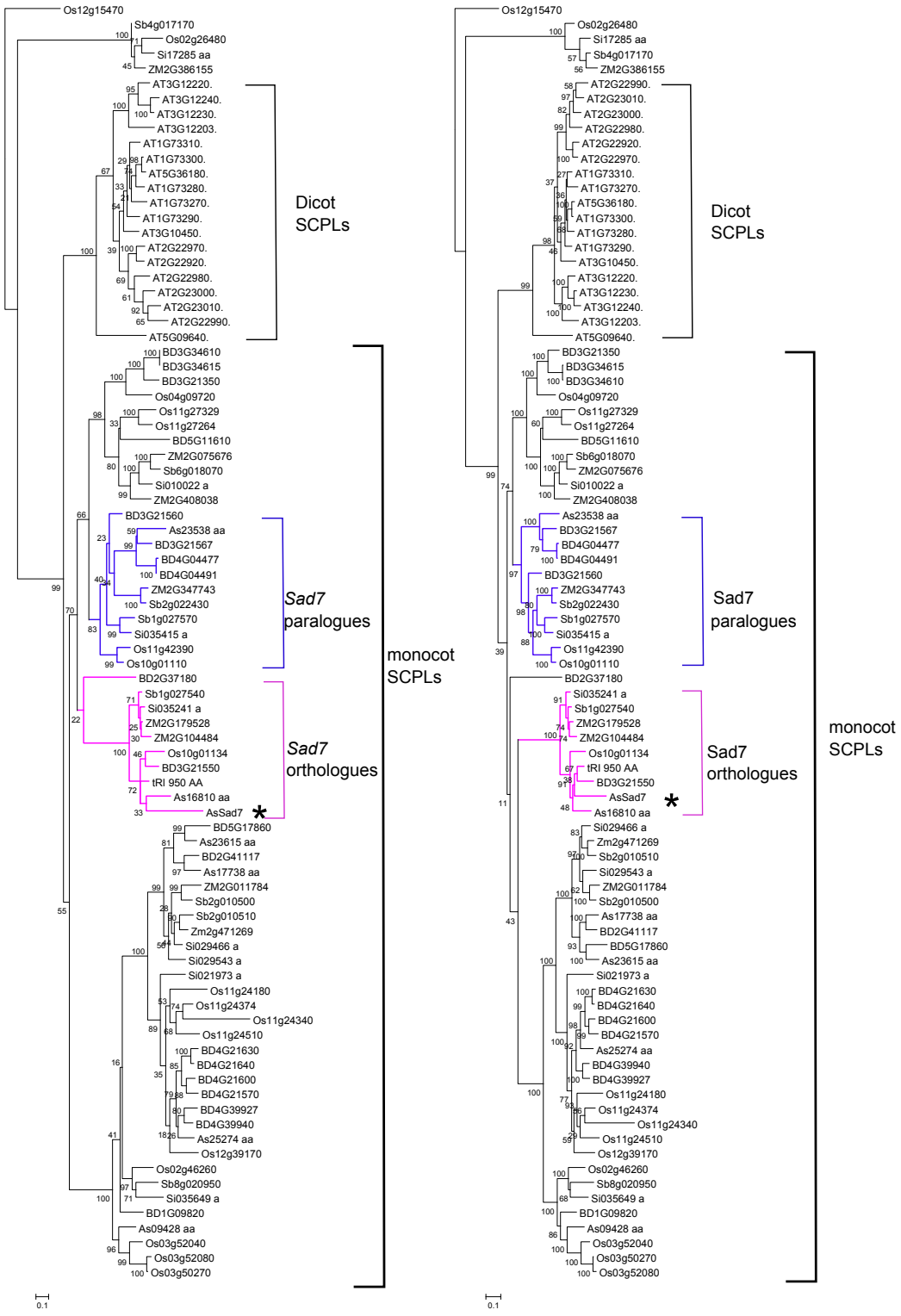


Figure 5.5: The bootstrapped maximum likelihood trees generated in RAxML 7.0.4 using the a) amino acid and b) coding sequence alignments of 85 Clade1A SCPLs. *Sad7* (*) is indicated by an asterisk. The numbers indicate the percentage of bootstrap replicates (out of 10,000) in which the given branching was observed. Raw tree files are in the Appendix (Trees 5.4 and 5.5)

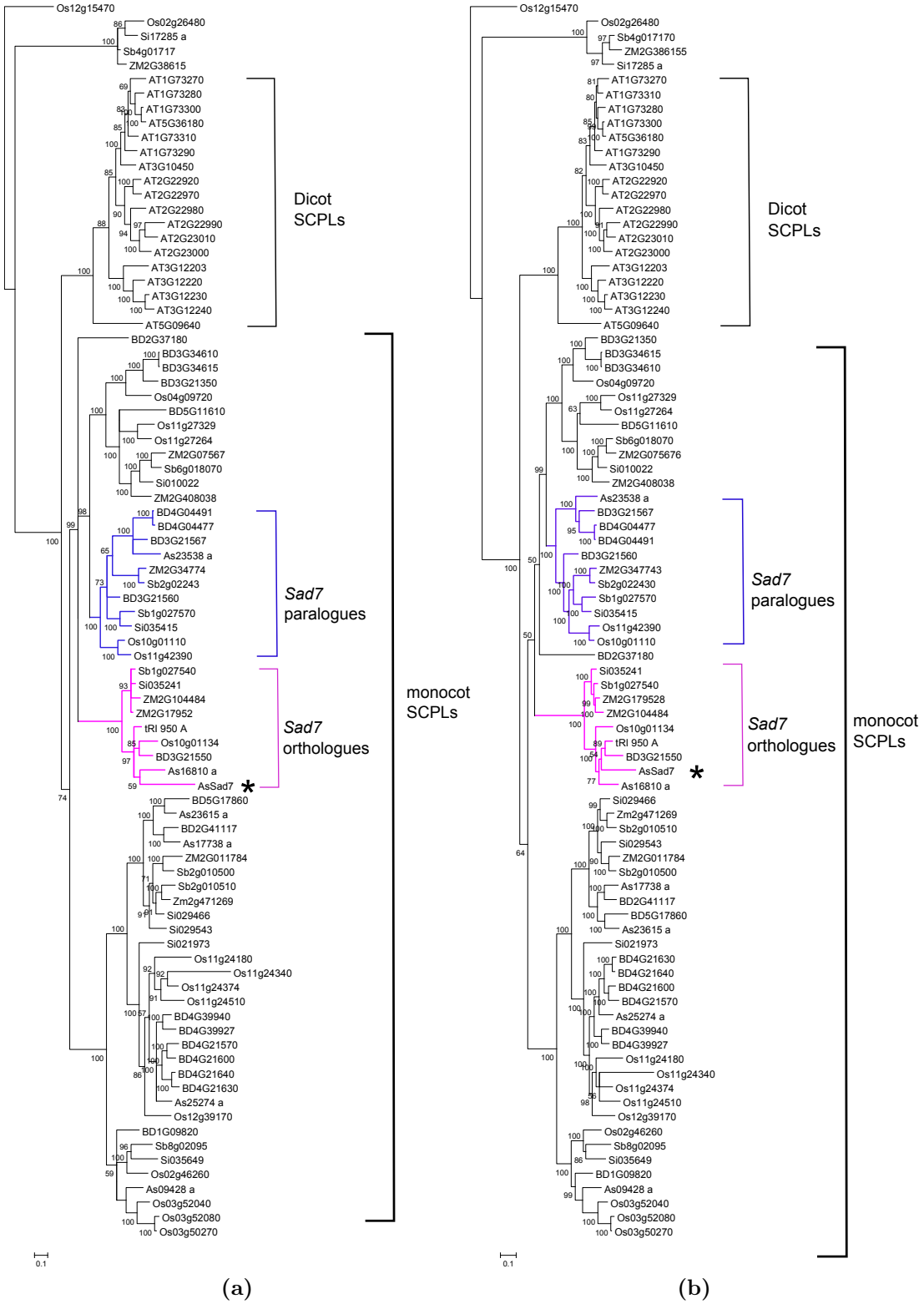
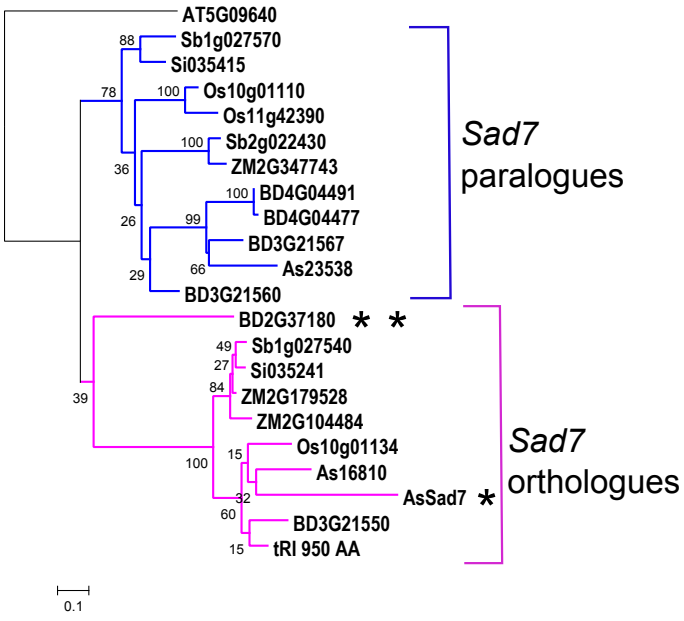
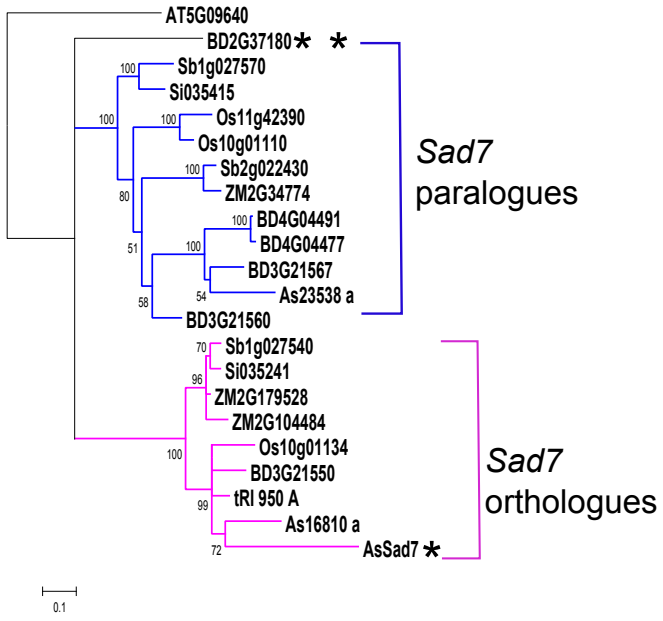


Figure 5.6: The 50 majority consensus trees generated in MrBayes 3.2.1 using the a) amino acid and b) coding sequence alignments of 85 Clade1A SCPLs. *Sad7* (*) is indicated by an asterisk. The bootstrap support of branching patterns (from 75,000,000 MCMC samples) are indicated. Raw tree files are in the Appendix (Trees 5.6 and 5.7)



(a)



(b)

Figure 5.7: Phylogenetic tree estimation of the *Sad7* orthologue and *Sad7* tandem paralogue groups only. a) tree constructed in RAxML 7.0.4. lnL = -5048.84, $\alpha = 1.01$ b) tree constructed in MrBayes 3.2.1. lnL = -5031.05, $\alpha = 0.91$. *Sad7* (*) and BRADI2G37180 (* *) are indicated on both trees. The bootstrap support of bifurcations (out of 10,000 or 75,000,000 replicates in RAxML and MrBayes respectively) are indicated. Raw tree files are in the Appendix (Trees 5.8 and 5.9)

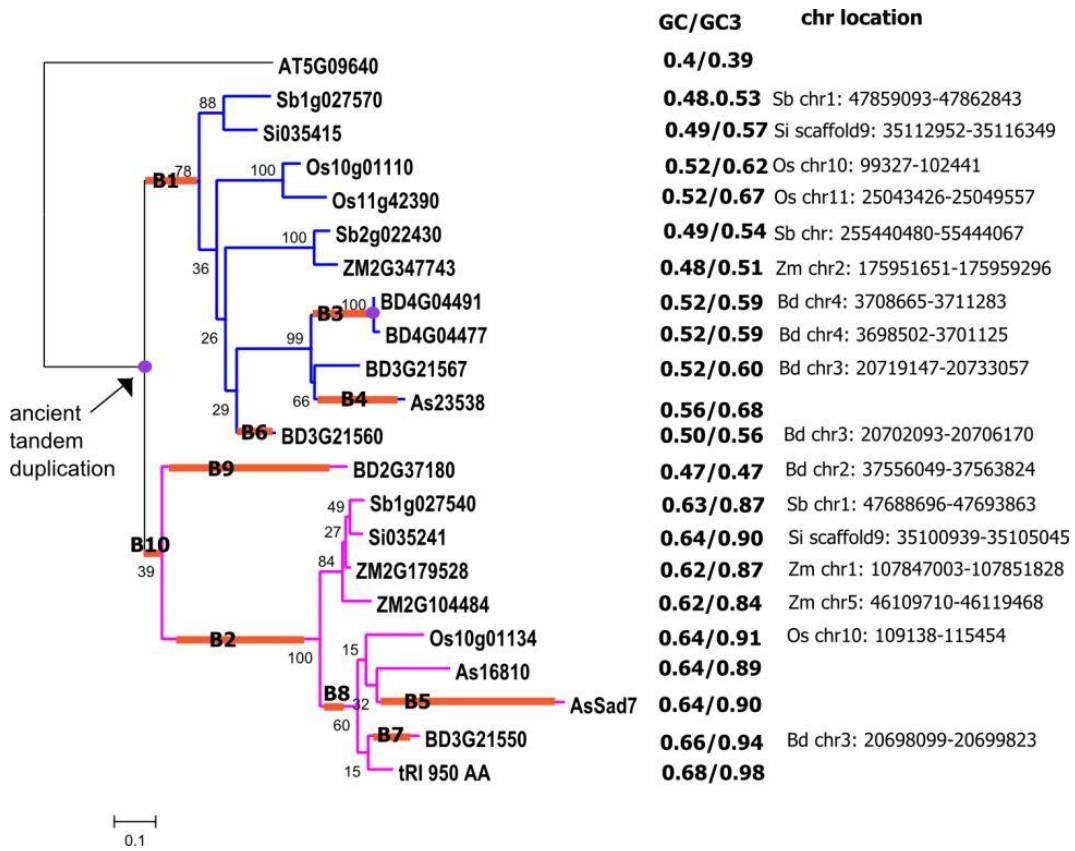


Figure 5.8: The RAxML amino acid *Sad7* orthologue and tandem paralogue tree (Figure 5.7a) with the ten branches selected for branch-site tests (B1-B10) labelled. Only B1 was found to be under positive selection. GC/GC_3 and chromosomal co-ordinates are listed next to each gene sequence.

that branch B1 possessed a significant signal of positive selection after Bonferroni correction for multiple testing, while the rest of the branches are under purifying selection. Two key amino acid residues (199 P (P = 0.993) and 211A (P = 0.992), both within conserved regions) were found to separate the *Sad7* orthologues from their tandem paralogues in the multiple sequence alignment (detailed branch-site test results are shown in Appendix Table 5.3). Additional features of the divergence of *Sad7* from its closely related homologues were discovered, including extensive loss of introns, elevated GC content and high levels of synonymous substitutions (dS), all in sequences subsequent to branch B2 (Appendix Table 5.4).

5.4.Results of the *Sad9* phylogenetic analysis

S-adenosyl-L-methionin (SAM) dependent *O*-methyltransferases (OMTs) catalyse the methylation of lignins, flavonoids, phytoalexins, and volatile compounds responsible for floral scent (Joshi and Chiang, 1998; Liscombe et al., 2012). The chemical mechanisms of methyl transfer reactions are identical among all plant OMTs, but the OMTs differ in their substrate selectivities. Comparative analysis of plant OMT cDNAs showed that they share high sequence identity (92-100%) within the last third of the protein sequence. It has been proposed that OMTs are separated into four main classes: class "A" OMTs methylate phenylpropanoid compounds; class "B" OMTs methylate flavonoid compounds; class "C" OMTs methylate alkaloids; and class "D" OMTs methylate aliphatic methyl acceptors (Barakat et al., 2011; Liscombe et al., 2012). OMT proteins that methylate the carboxyl group of various acids are classified into a fifth class (Liscombe et al., 2012). It has been reported that OMTs have probably originated from duplication of a gene encoding caffeic acid OMT (COMT) (Barakat et al., 2011). Phylogenetic studies indicated that tandem and segmental duplications have played a major role in the expansion of the OMT gene family in *Populus* (Barakat et al., 2011).

Sad9 encodes an S-adenosyl-L-methionine-dependent OMT involved in avenacin biosynthesis in oat, *Avena strigosa* (Mugford et al., 2013). SAD9 converts anthranilic acid to *N*-metyl anthranilate, which is then further processed by SAD10 to give the acyl donor NMA-Glu for SAD7 (Figure 1.2a) (Mugford et al., 2013). Here, phylogenetic studies of *Sad9* and related terpenoid methyltransferases in monocots are carried out to investigate how this class of OMTs has evolved.

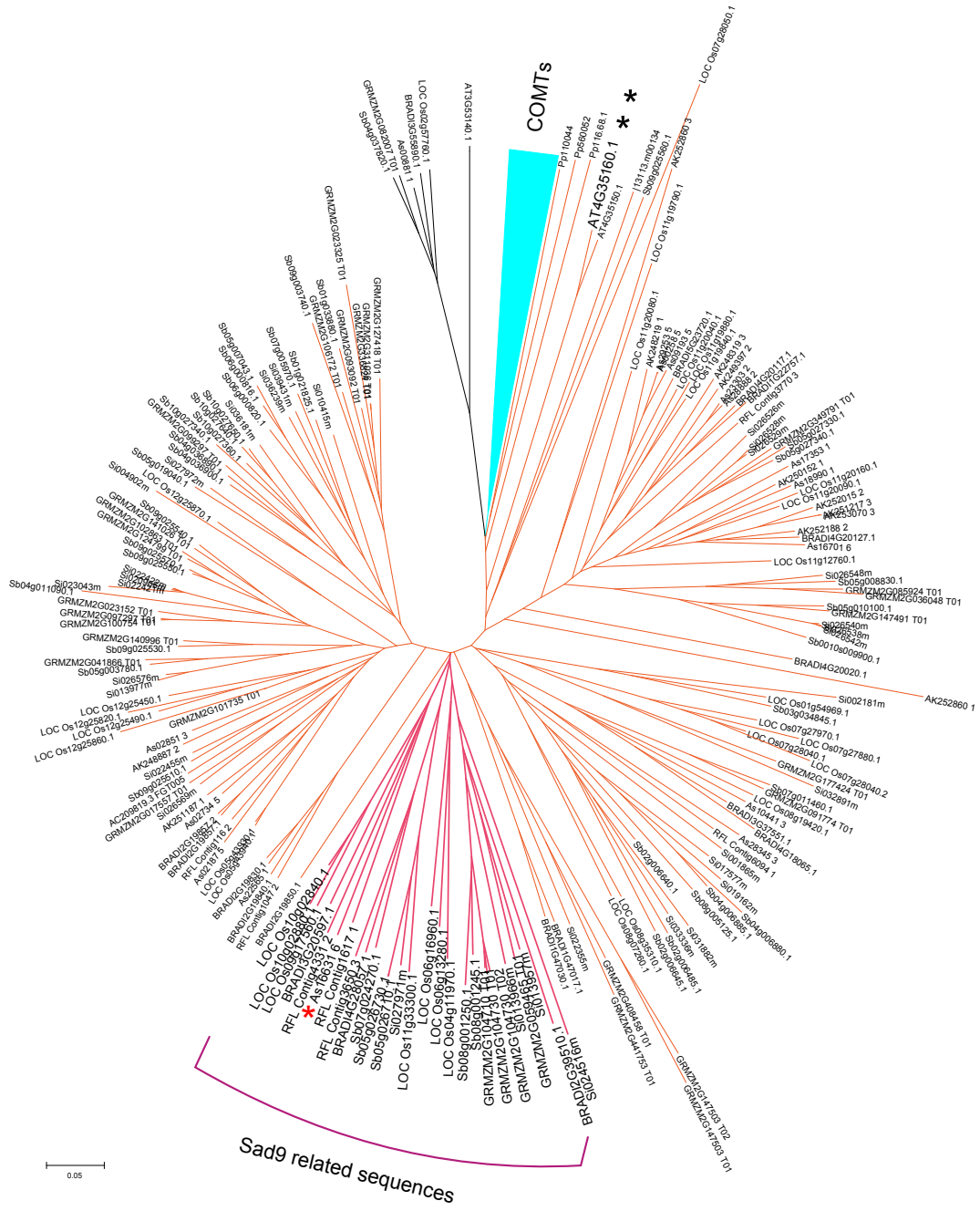


Figure 5.9: BioNJ tree of Sad9 homologues identified in the HMMsearch with E-value <1 x e^{-40}. With reference to the OMT tree from Barakat et al. (2011), the Class II OMT (COMT) clade (labelled in blue) is collapsed. Branches leading to the Class I OMT were labelled in orange and *Sad9*-related sequences in pink. *A. strigosa Sad9* (*) and AT4G35160 (* *) are indicated with aterisks. The original tree file may be found in the Appendix (Tree 5.10).

5.4.1. Sequence retrieval and annotation of *Sad9* homologues

The closest homologues of *A. strigosa Sad9* were identified in *A. thaliana, B. distachyon, S. bicolor*, and *Z. mays* using BLASTp (Gish, 1994). The sequences of the four genes were annotated, used to build a *Sad9*_closest_homologues HMM profile, and used to identify a total of 268 hits in the surveyed species: 20 in *A. thaliana*, 23 in *B. distachyon*, 47 in *O. sativa*, 3 in *P. patens*, 52 in *S. bicolor*, 41 in *S. italica*, 39 in *Z. mays*, 8 in wheat, 18 in barley, and 17 in oat. The resulting BioNJ tree estimated from the amino acid sequences of these hits (Figure 5.9) was then compared to the reference phylogenetic tree of *O*-methyltransferases (Barakat et al., 2011). Plant OMTs were designated into Class I and Class II in the reference tree. The closest homologues of *Sad9* in *O. sativa* and *S. bicolor* are both located in Class I. In this particular Class I clade, no *A. thaliana* nor *P. patens* genes were identified, suggesting that this group of methyltransferases are monocot specific.

A subset of 175 Class I OMTs identified in the BioNJ tree were selected for gene annotation and phylogenetic analysis. *AthOMT12* (AT4G35160), which was indicated to be the closest homologue outside the Class I OMT clade in the BioNJ tree (Figure 5.9), was designated as the outgroup for later phylogenetic tree estimation. The 157 monocot Class I OMT sequences were manually annotated while the two *A. thaliana* sequences were directly downloaded from TAIR10 (Lamesch et al., 2012) and Phytozome v8.0 (Goodstein et al., 2012) respectively. The locus name, genomic orientation, reference cDNA accession number and gene structure description are listed in Appendix Table 5.5. The reconstructed sequences of each locus were listed in Appendix Table 5.6 with information on insertions, deletions, or in-frame stop codons indicated.

5.4.2. Multiple sequence alignment and phylogenetic tree estimation of Class I OMT sequences

After annotation, 157 reconstructed amino acid sequences were aligned. A coding sequence alignment was made using Pal2Nal (Suyama et al., 2006). A total of 25 sequences were removed from the alignment because they were either truncated or highly divergent from the rest of the Class I OMTs. The selected evolutionary models for these alignments are shown in Table 5.3. The overall topologies of the four resulting phylogenetic trees were found to be highly similar to one another (Figures 5.10 and 5.11).

The Class I OMT trees contain three main groups (groups 1-3 in Figures 5.10 and

Summary of Class I OMT phylogenetic trees						
Phylogenetic software	Alignment type	Number of sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	134	275 aa	-31301.57	1.55	JTT + I + Γ + F
RAxML	codon	134	822 bp	-58496.91	1.07	GTR + Γ
MrBayes	amino acid	134	275 aa	-31135.59	1.45	JTT + I + Γ + F
MrBayes	codon	134	822 bp	-58398.31	1.08	GTR + Γ

Table 5.3: Summary of Class I OMT phylogenetic analyses

5.11), with one of the clades (group 3) comprising the closest related homologues of *Sad9*. The remaining two major clades (groups 1 and 2) each contain several tandem arrays of OMTs, with a *Panicoideae*-specific clade in group 2 (clade P in Figure 5.10 and 5.11). For example, seven tandemly duplicated OMTs on rice chromosome 11 (highlighted in red) and three tandem duplicates on *B. distachyon* chromosome 4 (highlighted in green) are located within group 1. Amongst the group 2 OMT sequences, a tandem array on rice chromosome 12 (highlighted in blue) was found to be closely related to two tandem duplicate OMTs on rice chromosome 5, likely to be lineage specific expansions due to a paucity of closely related sequences in *B. distachyon*, *A. strigosa* and wheat. Tandem arrays were also found on *S. bicolor* chromosome 9 (highlighted in light green), but their orthologues were scattered across three chromosomal regions in *Z. mays* and *S. italica*.

The topology of the clade containing *Sad9*-related sequences (group 3 in Figures 5.10 and 5.11) differed slightly between the amino acid and coding sequence trees. In order to improve the resolution of this group, an additional phylogenetic analysis was carried out using only the closest homologues of *Sad9* (the 23 sequences in group 3, highlighted in pink in Figure 5.10), with As22565 and BRADI2G19830 included as outgroup sequences. The resulting topologies of the *Sad9* orthologue trees are again highly similar (Figures 5.12 and 5.13), but still with minor differences between the amino acid and coding sequence trees. Phylogenetic estimation revealed that *A. strigosa* S75 *Sad9* possesses a very closely related homologue in wheat, Tri4331. No sequences from *S. bicolor, S. italica*, and *Z. mays* grouped with AsMT1, suggesting that AsMT1 arose after the divergence of the BEP ancestor from the other *Poaceae*, or that the corresponding ancestral *AsMT1* orthologue was lost in the *Panicoideae*.

5.4.3.Molecular evolution analysis of *Sad9* homologues

Branch-sites tests were performed on five branches of interest in the *Sad9* orthologue tree (Figures 5.12a and 5.14). Only branch B1 showed a significant signal of positive selection, while the rest of the branches are under purifying selection. The results of the branch-site tests are summarized in Appendix Table 5.7. The GC contents of the 25 coding sequences were measured (Appendix Table 5.8). Values for the closest related homologues of *Sad9* were relatively similar (Figure 5.14), with an overall GC content of between 0.53 and 0.60, and a GC_3 content ranging from 0.63 to 0.90. Unlike the *Sad7* orthologous tree, no GC content enrichment of any particular clade was observed. Interestingly, the GC contents of the outgroup sequences As22565 (GC/GC_3: 0.45/0.48) and BRADI2G19830 (GC/GC_3: 0.45/0.49) were found to be significantly lower than the remainder of the tree.

5.5.Results of the *Sad10* phylogenetic analysis

Sad10 (UGT74H5) (Owatworakit et al., 2013) is a member of the family 1 uridine diphosphate-dependent glycosyltransferases (UGTs), which are involved in glycosylating terpenoids, benzoates, flavonoids, phenylpropanoids and plant hormones. Family 1 UGTs generally catalyse transfer of the glycosyl group from UDP-activated sugars to acceptor molecules. The family 1 UGTs are characterized by a signature plant secondary product glycosyltransferase (PSPG) box motif for binding UDP-sugar (Gachon et al., 2005). In the avenacin biosynthetic pathway, SAD10 attaches a glycosyl group to an N-methyl anthranilic acid that is then attached to the β-amyrin backbone by SAD7 (Figure 1.2a) (Mugford et al., 2009; Owatworakit et al., 2013). Genome-wide comparative analysis of family 1 UGTs has previously been carried out in the sequenced genomes of *A. thaliana*, algae, mosses, dicots, monocots and trees (Caputi et al., 2012). UGT family 1 enzymes were classified into phylogenetic groups A to N. SAD10 belongs to the group L UGTs, which consist of UGT enzymes belonging to families 74, 75 and 84.

5.5.1.Sequence retrieval and annotation of *Sad10* homologues

A BLASTp (Gish, 1994) search identified the closest hits to the amino acid sequence of *A. strigosa Sad10* in *B. distachyon, O. sativa, S. bicolor*, and *Z. mays* respectively. An HMM profile was constructed from the annotated sequences of

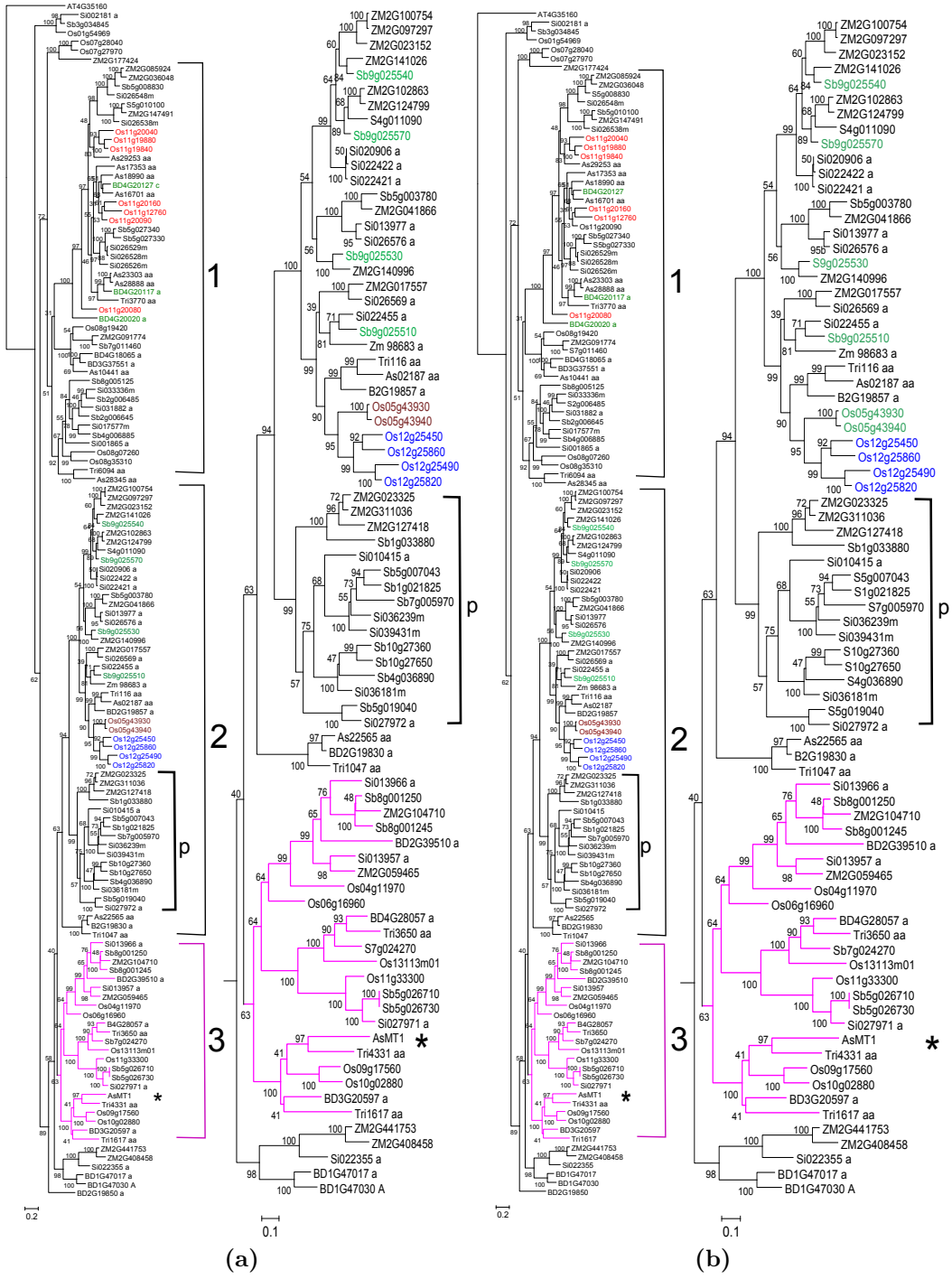


Figure 5.10: The bootstrapped maximum likelihood trees of Class I OMT genes estimated in RAxML 7.0.4 using the a) amino acid and b) coding sequence alignments. Left: The complete OMT tree. Right: Loci closely related to *Sad9* (groups 2 and 3). The *Panicoideae*-specific clade (labelled p) is indicated. *A. strigosa Sad9* (encoding the enzyme AsMT1) is indicated with an asterisk. Tandem arrays of genes are colour coded. The numbers indicate the percentage of bootstrap replicates (out of 10,000) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.11 and 5.12).

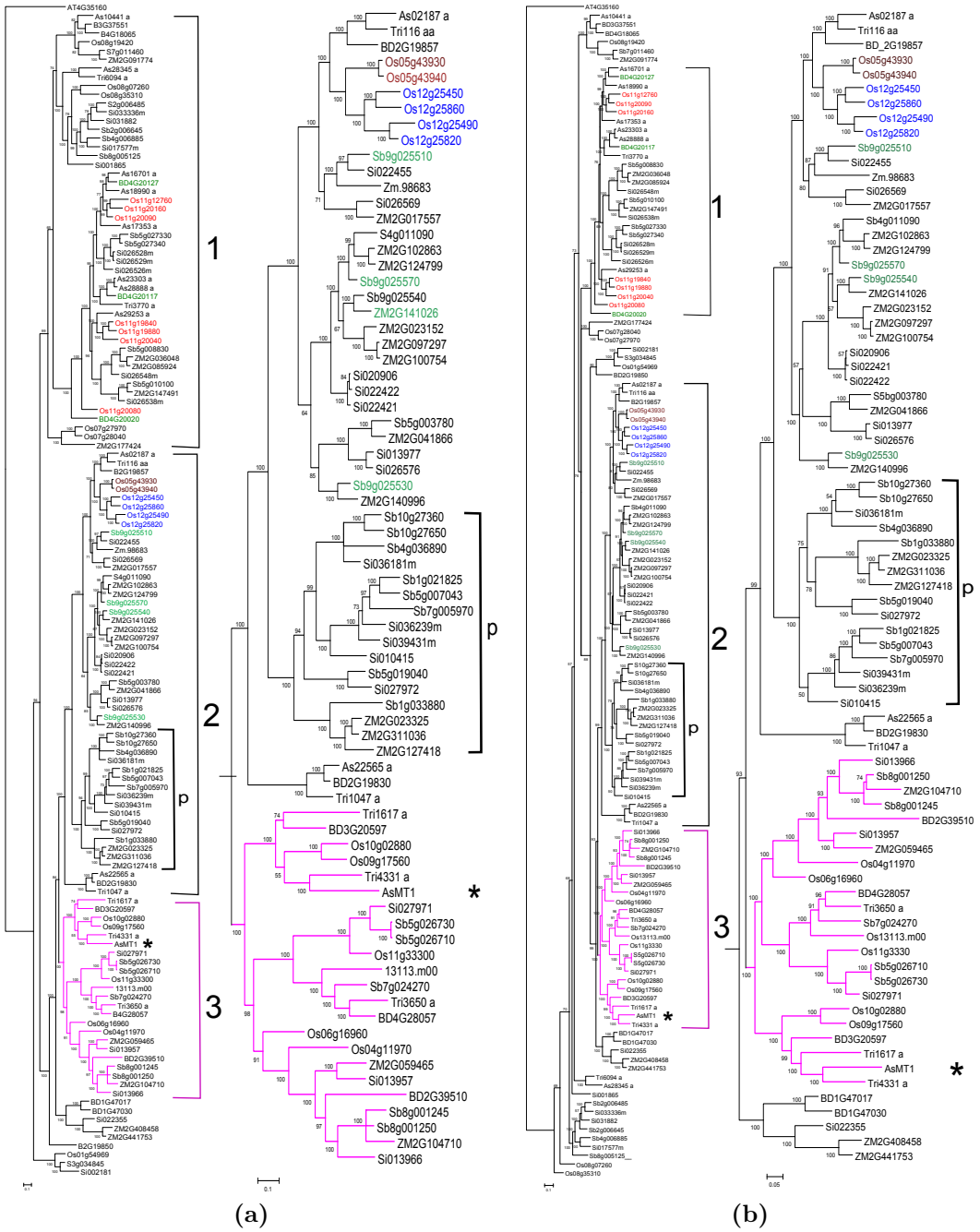
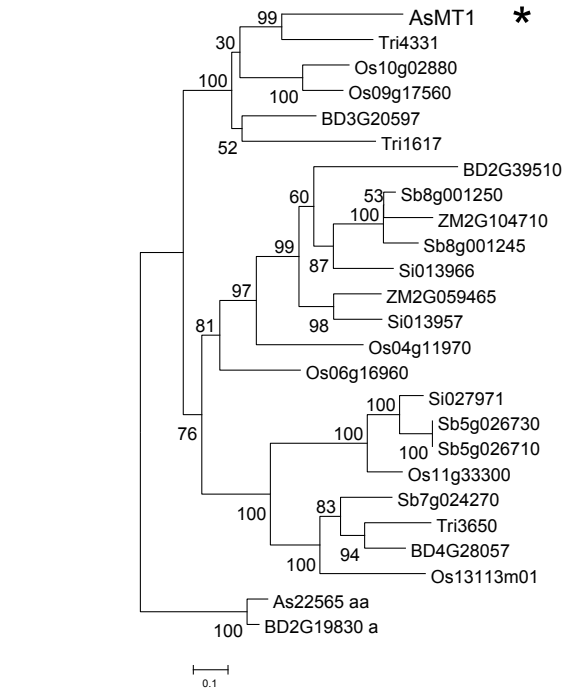
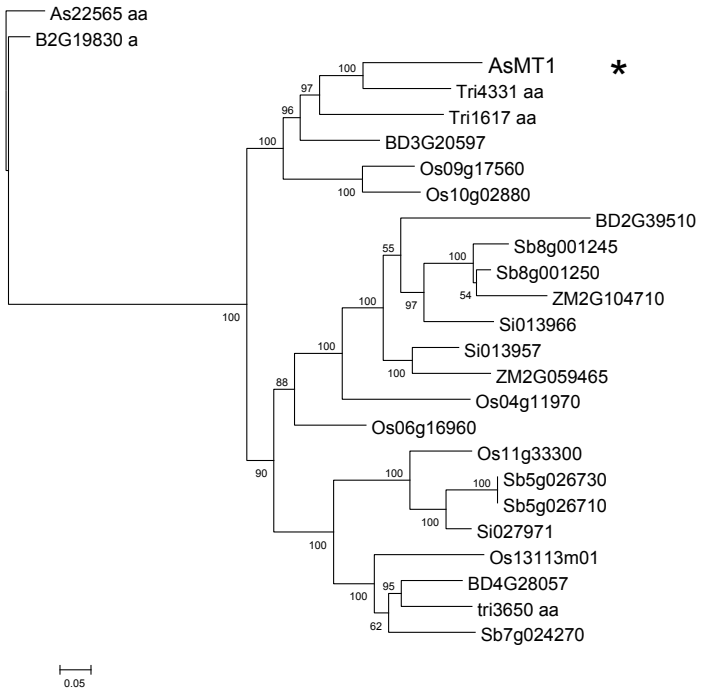


Figure 5.11: The 50 majority consensus trees generated in MrBayes 3.2.1 using the a) amino acid and b) coding sequence alignments for Class I OMT sequences. Left: The complete OMT tree. Right: Loci closely related to *Sad9* (groups 2 and 3). *A. strigosa Sad9* (encoding the enzyme AsMT1) is indicated with an asterisk. The numbers indicate the percentage of bootstrap replicates (out of 75,000.000 MCMC samples) in which the given branching was observed. Tandem arrays of genes are colour coded. Original tree files are in the Appendix (Trees 5.13 and 5.14).

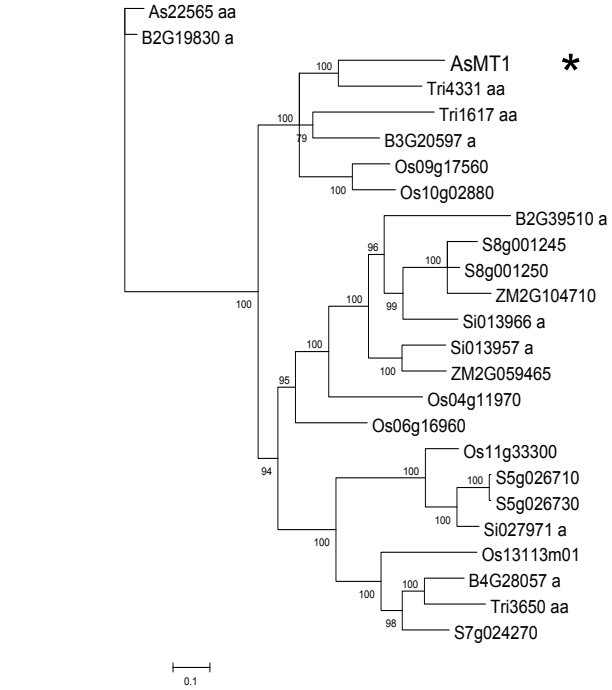


(a)

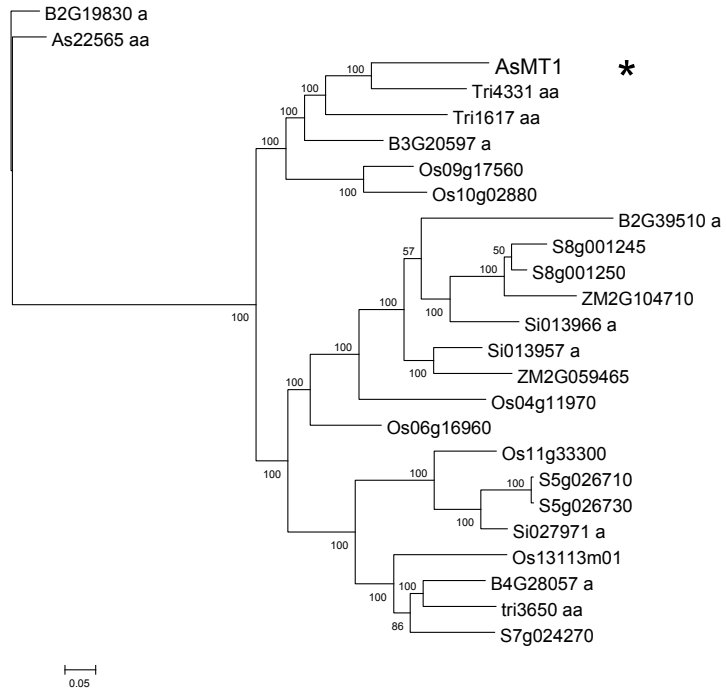


(b)

Figure 5.12: Phylogenetic trees estimated in RAxML 7.0.4 using (a) the amino acid alignment (lnL = -7077.84, α = 1.43) and (b) the coding sequence alignment (lnL = -12795.53, α = 1.06) of the *Sad9* orthologues. *Sad9* AsMT1 is indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 10,000) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.15 and 5.16).



(a)



(b)

Figure 5.13: Phylogenetic trees estimated in MrBayes 3.2.1 using (a) the amino acid alignment (lnL = -7135.94, $\alpha = 1.63$) and (b) the coding sequence alignment (lnL = -12853.48, $\alpha = 1.06$) of the *Sad9* orthologues. *Sad9* AsMT1 is indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 75,000,000 MCMC samples) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.17 and 5.18).

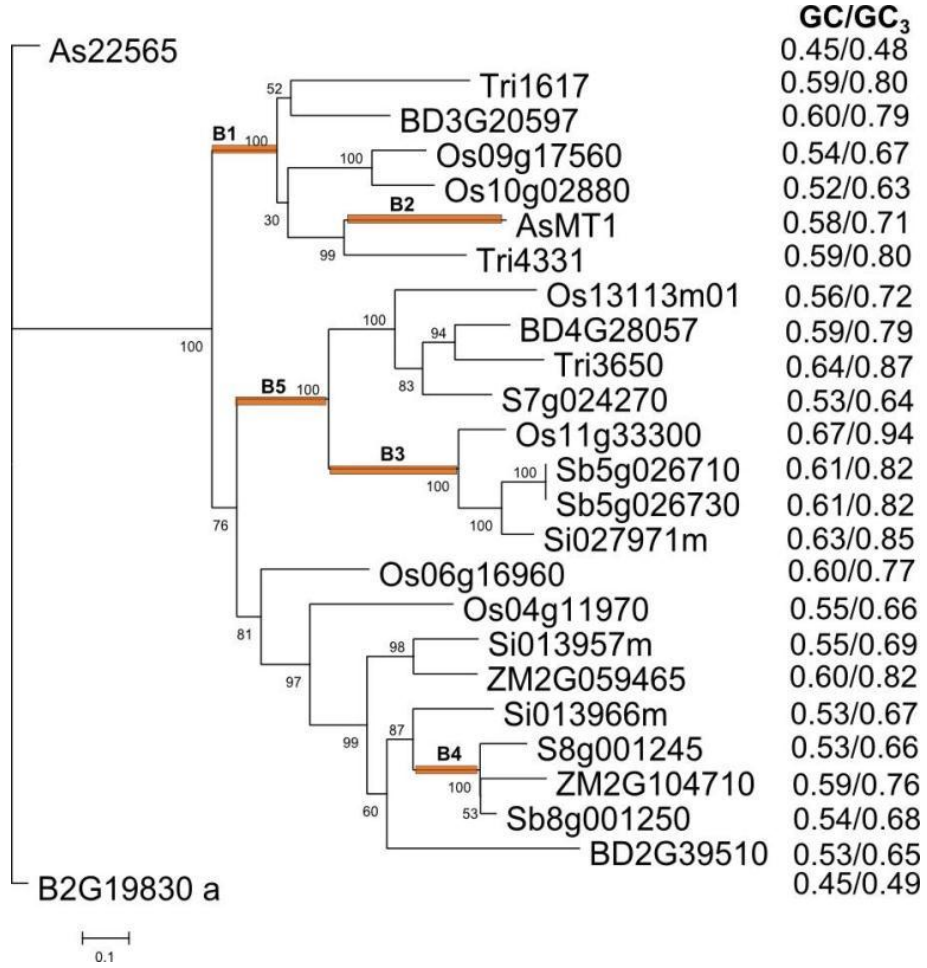


Figure 5.14: Branch-site tests performed on the *Sad9* orthologue tree estimated in RAxML 7.0.4 from the amino acid alignment. GC/GC_3 content of each gene is listed next to the gene sequence in the tree.

these hits and used to uncover over 600 sequences in the surveyed species (15 barley, 172 *B. distachyon*, 135 oat, 24 *P. patens*, 224 *O. sativa*, 227 *S. italica*, 220 *S. bicolor*, 24 wheat and 215 *Z. mays* sequences). The HMMsearch results were consistent with the expected numbers of putative UGTs from each plant genome surveyed (Caputi et al., 2012). A BioNJ tree was constructed from 567 full length hits with an E-value $<1 \times e^{-50}$ (Appendix Tree 5.19).
The 113 Group L glycosyltransferases, to which *Sad10* belongs, were identified from the BioNJ tree with reference to the phylogenetic trees of UGT74s, UGT75s and UGT84s (Owatworakit et al., 2013) and were annotated (Appendix Table 5.19). Only five of the 113 annotated sequences (BRADI1G37070, BRADI4G42997, LOC_Os04g12720, GRMZM2G319965, and Sb10g023050) were predicted to contain frame-shift deletions or in-frame stop codons, and therefore likely to encode non-functional proteins. Details of gene structure, coding and amino acid sequences for each annotated Group L UGT member were recorded (Appendix Tables 4.9 and 4.10).
Sequences from the UGT74 family were found to contain greater numbers of exons (two to five) than UGT75 or UGT84 family sequences (one to three). In UGT74 family sequences, conservation of exon junctions was noted both in the presence and absence of introns. It could therefore be concluded that the ancestral gene structure of UGT74s likely contained three exons, and that further genomic insertions/deletions have led to the changes in gene structures to give one, two and five exons.

5.5.2.Multiple sequence alignment and phylogenetic tree estimation of Group L UGTs

Manual refinement of the multiple sequence alignments resulted in the removal of six sequences (truncated sequences: Pp1s77_100V6, Si008274, LOC_Os1g49230, Sb05g002700, and BRADI1G11940; divergent sequence: GRMZM2G319965). The selected evolutionary models for these alignments are shown in Table 5.4. Phylogenetic trees were rooted using the outgroup sequences Pp1s4_21V6 and Pp1s6_30V6, located within a neighboring clade of the Group L UGTs. The Group L UGTs were clearly split into three clades in the resulting RAxML and MrBayes phylogenetic trees (Figures 5.15 and 5.16; Table 5.4), which were functionally classified as UGT75s, UGT84s and UGT74s. However, only RAxML and MrBayes coding sequence trees were found to possess topologies that separated the *A. thaliana* UGTs from the rest of the monocot genes. This gene tree incongruity suggested that although the dicot UGTs and monocot UGTs shared amino acid sequence similarities, their coding sequences

Summary of Group L UGT phylogenetic trees						
Phylogenetic software	Alignment type	Number of sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	105	247 aa	-22615.24	1.40	JTT + I + Γ
RAxML	codon	105	741 bp	-44110.34	1.05	GTR + Γ
RAxML	third_base_strip	105	494	-22917.17	1.12	GTR + Γ
MrBayes	amino acid	105	247 aa	-22849.43	1.50	JTT + I + Γ
MrBayes	codon	105	741 bp	-44227.82	1.10	GTR + Γ
MrBayes	third_base_strip	105	494	-23044.36	1.12	GTR + Γ

Table 5.4: Summary of Group L UGT phylogenetic analyses

differed more markedly.

To assess the impact of third base synonymous changes on the tree structure, a phylogenetic analysis of the third_base_strip coding sequence was carried out. The resulting phylogenetic trees possessed topologies consistent with those derived from the amino acid alignment (Figures 5.15a,c and 5.16a,c), confirming the role of third base changes in the previous topology inconsistency. Interestingly, *Sad10* was shown to group with BRADI4G35350, Sb02g030020, LOC_OS09g34214 and Si29697m in the third_base_strip trees (Figures 5.15c and 5.16c) with high bootstap support, different from the amino acid and full coding sequence trees (Figures 5.15a,b and 5.16a,b), casting some doubt on the identities of its closest homologues. The altered grouping suggests either that long-branch attraction of *Sad10* may have occured in the third_base_strip tree estimation or that *Sad10* is indeed closely related to these genes.

A phylogenetic analysis of *Sad10* and its closest homologues (23 monocot UGT74 sequences) was carried out, using LOC_Os03g48740 as an outgroup sequence. The resulting trees exhibited topologies with higher bootstrap support (Figure 5.17) than in the broader UGT analysis. Furthermore, three of the four trees possessed highly consistent groupings, the exception being the MrBayes coding sequence tree (Figure 5.17d). All four trees indicated that *Sad10* is distantly related to the rest of the monocot UGT74 sequences. On the contrary, the oat cDNA contig As07784 grouped closely with the *B. distachyon* UGT74s BRADI5G03330 and BRADI5G03390 (Figure 5.17). Tandem arrays of UGT74s were discovered on *B. distachyon* chromosome 5 (blue) and *O. sativa* chromosome 4 (green) (Figure 5.17).

5.5.3.Molecular evolution analysis of *Sad10* homologues

Branch-site tests were carried out on three branches of the RAxML amino acid *Sad10* homologues tree. Each selected branch was tested under the model M1a (neutral selection, foreground $\omega = 1$) and M2a (positive selection, foreground ω <1) (Figure 5.18) in PAML 4.5. All three of the branches tested were shown to have evolved under purifying selection. Details of the branch-site test results are given in Appendix Table 5.11.

The GC and GC_3 contents of all *Sad10* homologues are relatively high, a signature of many monocot genes (Figure 5.18). However, the clade containing Sb02g030020, Si029697, BRADI4G35350, and LOC_Os09g34214 was shown to possess elevated GC and GC_3 contents (GC_3> 0.90). These values may have had an effect on the branch-site tests performed on branches B1 and B2, inflating *dS* estimates as may also have occured in the *Sad7* selection tests.

Fairly large differences in GC and GC_3 contents were observed in the *B. distachyon* tandem array (Figure 5.18). It could be speculated that these tandemly arranged UGT74s might possess different gene expression patterns due to the altered GC_3 content (Tatarinova et al., 2010), which could be the first evolutionary step towards neo-functionalization or specialization to a particular metabolic pathway. A similar pattern of GC content differences was also observed in the *O. sativa* tandem array, but not for those in *S. bicolor* or *S. italica*.

5.6.Results of the *Sad2* phylogenetic analysis

Sad2 (which encodes the enzyme AsCYP51H10) originated from an ancient duplication event. This event led to the divergence of the CYP51H subfamily from the CYP51G subfamily, the cytochrome P450s involved in primary sterol biosynthesis (Qi et al., 2006). SAD2 is the first tailoring enzyme of avenacin biosynthesis, modifying the β-amyrin backbone (Figure 1.2a) (Geisler et al., 2013; Qi et al., 2006). In rice, a total of 12 *CYP51* genes have been previously identified (Inagaki et al., 2011). Inagaki et al. (2011) showed that CYP51H members were monocot-specific, similar to *Sad7* and *Sad9*. These observations collectively suggest that the avenacin pathway began to emerge after the monocot-dicot split. Here, the previous analyses of *Sad2* are extended, to shed further light on the evolution of the CYP51 gene family.

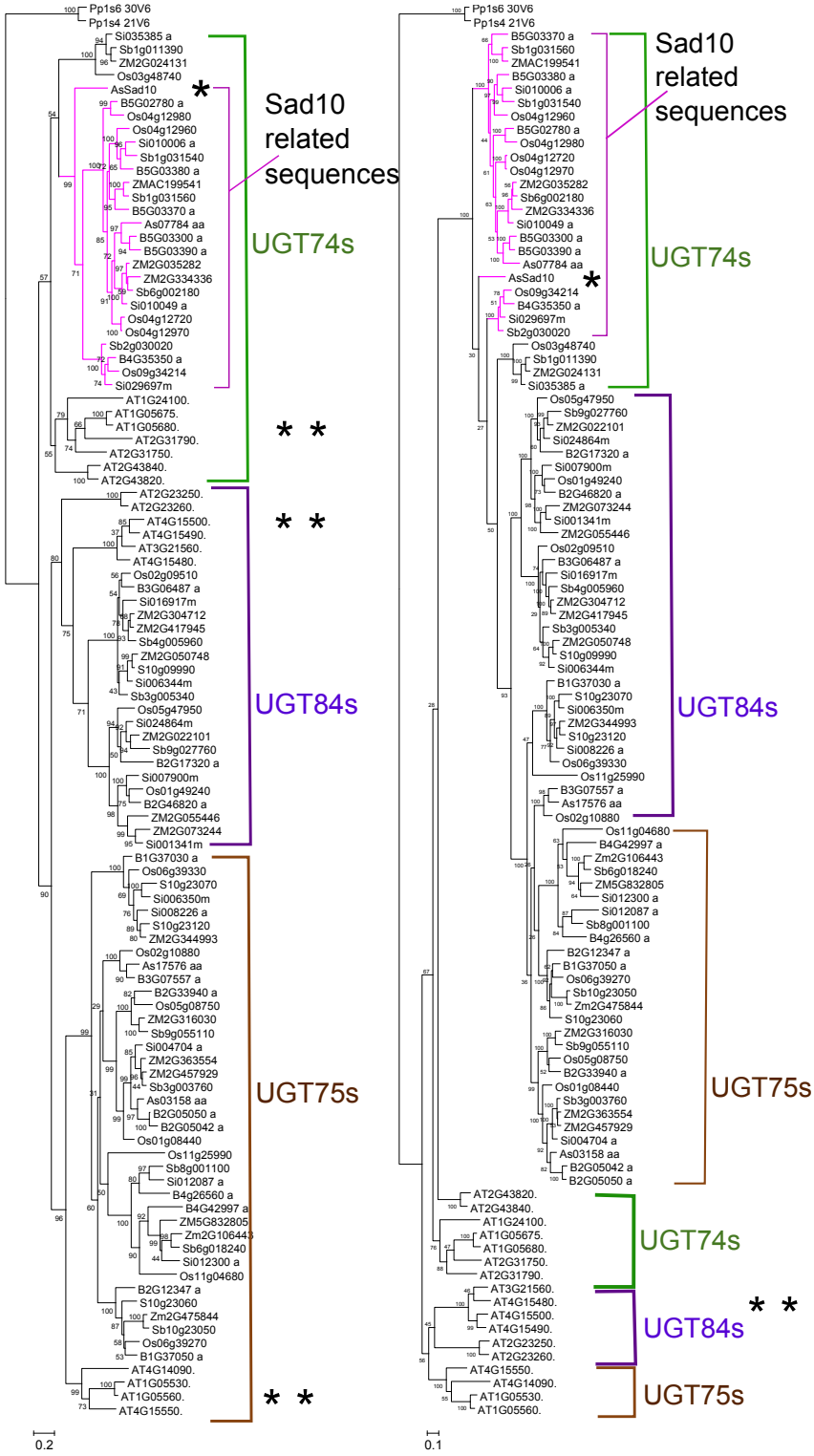
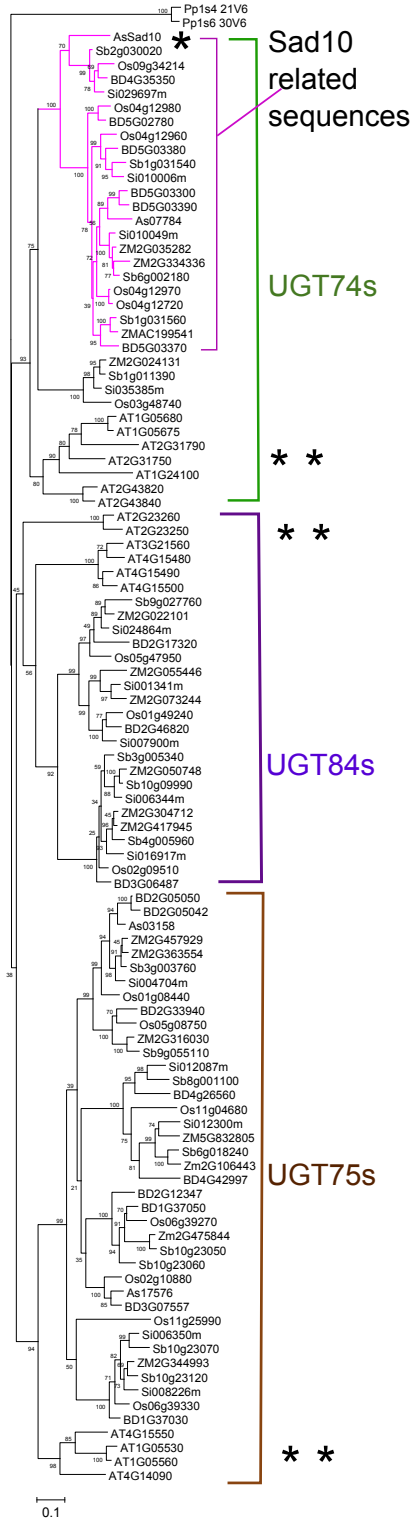


Figure 5.15: The bootstrapped maximum likelihood trees generated in RAxML 7.0.4 using the a) amino acid, b) coding sequence and c) third_base_strip alignments. The *A. strigosa Sad10* sequence (*) and the *A. thaliana* sequences (* *) are indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 10,000) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.20 to 5.22).


Pp1s4 21V6
Pp1s6 30V6
AsSad10
Sb2g030020
Os09g34214
BD4G35350
Si029697m
Os04g12980
BD5G02780
Os04g12960
BD5G03380
Sb1g031540
Si010006m
BD5G03300
BD5G03390
As07784
Si010049m
ZM2G035282
ZM2G334336
Sb6g002180
Os04g12970
Os04g12720
Sb1g031560
ZMAC199541
BD5G03370
ZM2G024131
Sb1g011390
Si035385m
Os03g48740
AT1G05680
AT1G05675
AT2G31790
AT2G31750
AT1G24100
AT2G43820
AT2G43840
AT2G23260
AT2G23250
AT3G21560
AT4G15480
AT4G15490
AT4G15500
Sb9g027760
ZM2G022101
Si024864m
BD2G17320
Os05g47950
ZM2G055446
Si001341m
ZM2G073244
Os01g49240
BD2G46820
Si007900m
Sb3g005340
ZM2G050748
Sb10g09990
Si006344m
ZM2G304712
ZM2G417945
Sb4g005960
Si016917m
Os02g09510
BD3G06487
BD2G05050
BD2G05042
As03158
ZM2G457929
ZM2G363554
Sb3g003760
Si004704m
Os01g08440
BD2G33940
Os05g08750
ZM2G316030
Sb9g055110
Si012087m
Sb8g001100
BD4g26560
Os11g04680
Si012300m
ZM5G832805
Sb6g018240
Zm2G106443
BD4G42997
BD2G12347
BD1G37050
Os06g39270
Zm2G475844
Sb10g23050
Sb10g23060
Os02g10880
As17576
BD3G07557
Os11g25990
Si006350m
Sb10g23070
ZM2G344993
Sb10g23120
Si008226m
Os06g39330
BD1G37030
AT4G15550
AT1G05530
AT1G05560
AT4G14090
0.1
Sad10 related sequences
UGT74s
UGT84s
UGT75s
*
* *
* *
* *

(c)

Figure 5.15

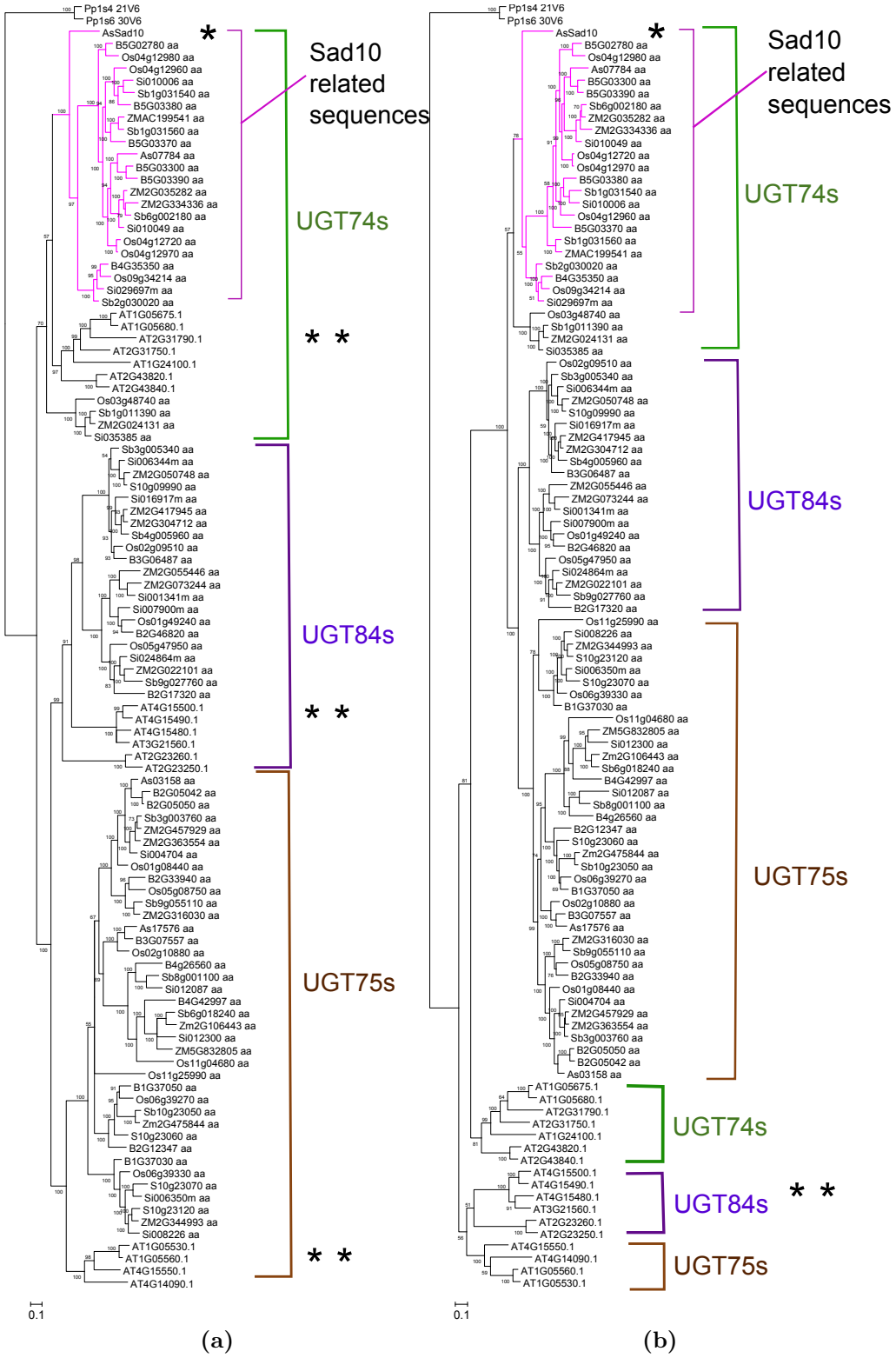
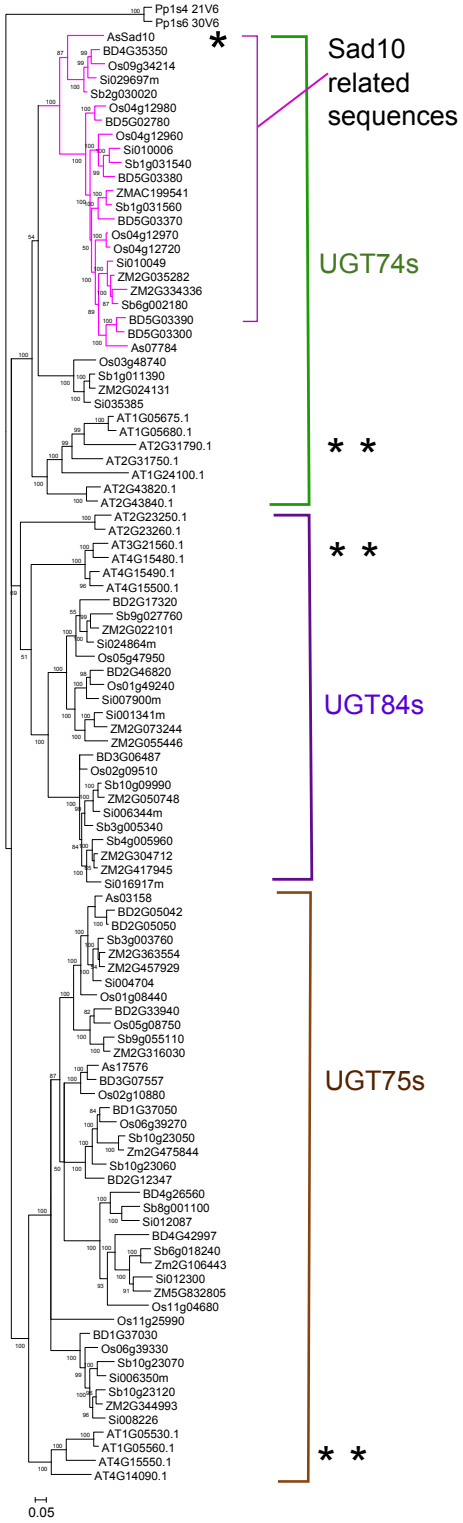


Figure 5.16: The 50 majority consensus trees generated in MrBayes 3.2.1 using the a) amino acid, b) coding sequence and c) third_base_strip alignments. The *A. strigosa Sad10* sequence (*) and the *A. thaliana* sequences (* *) are indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 75,000,000 MCMC samples) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.23 to 5.25).


Pp1s4 21V6
Pp1s6 30V6
AsSad10
BD4G35350
Os09g34214
Si029697m
Sb2g030020
Os04g12980
BD5G02780
Os04g12960
Si010006
Sb1g031540
BD5G03380
ZMAC199541
Sb1g031560
BD5G03370
Os04g12970
Os04g12720
Si010049
ZM2G035282
ZM2G334336
Sb6g002180
BD5G03390
BD5G03300
As07784
Os03g48740
Sb1g011390
ZM2G024131
Si035385
AT1G05675.1
AT1G05680.1
AT2G31790.1
AT2G31750.1
AT1G24100.1
AT2G43820.1
AT2G43840.1
AT2G23250.1
AT2G23260.1
AT3G21560.1
AT4G15480.1
AT4G15490.1
AT4G15500.1
BD2G17320
Sb9g027760
ZM2G022101
Si024864m
Os05g47950
BD2G46820
Os01g49240
Si007900m
Si001341m
ZM2G073244
ZM2G055446
BD3G06487
Os02g09510
Sb10g09990
ZM2G050748
Si006344m
Sb3g005340
Sb4g005960
ZM2G304712
ZM2G417945
Si016917m
As03158
BD2G05042
BD2G05050
Sb3g003760
ZM2G363554
ZM2G457929
Si004704
Os01g08440
BD2G33940
Os05g08750
Sb9g055110
ZM2G316030
As17576
BD3G07557
Os02g10880
BD1G37050
Os06g39270
Sb10g23050
Zm2G475844
Sb10g23060
BD2G12347
BD4g26560
Sb8g001100
Si012087
BD4G42997
Sb6g018240
Zm2G106443
Si012300
ZM5G832805
Os11g04680
Os11g25990
BD1G37030
Os06g39330
Sb10g23070
Si006350m
Sb10g23120
ZM2G344993
Si008226
AT1G05530.1
AT1G05560.1
AT4G15550.1
AT4G14090.1
0.05
Sad10 related sequences
UGT74s
UGT84s
UGT75s
*
* *
* *
* *

(c)

Figure 5.16

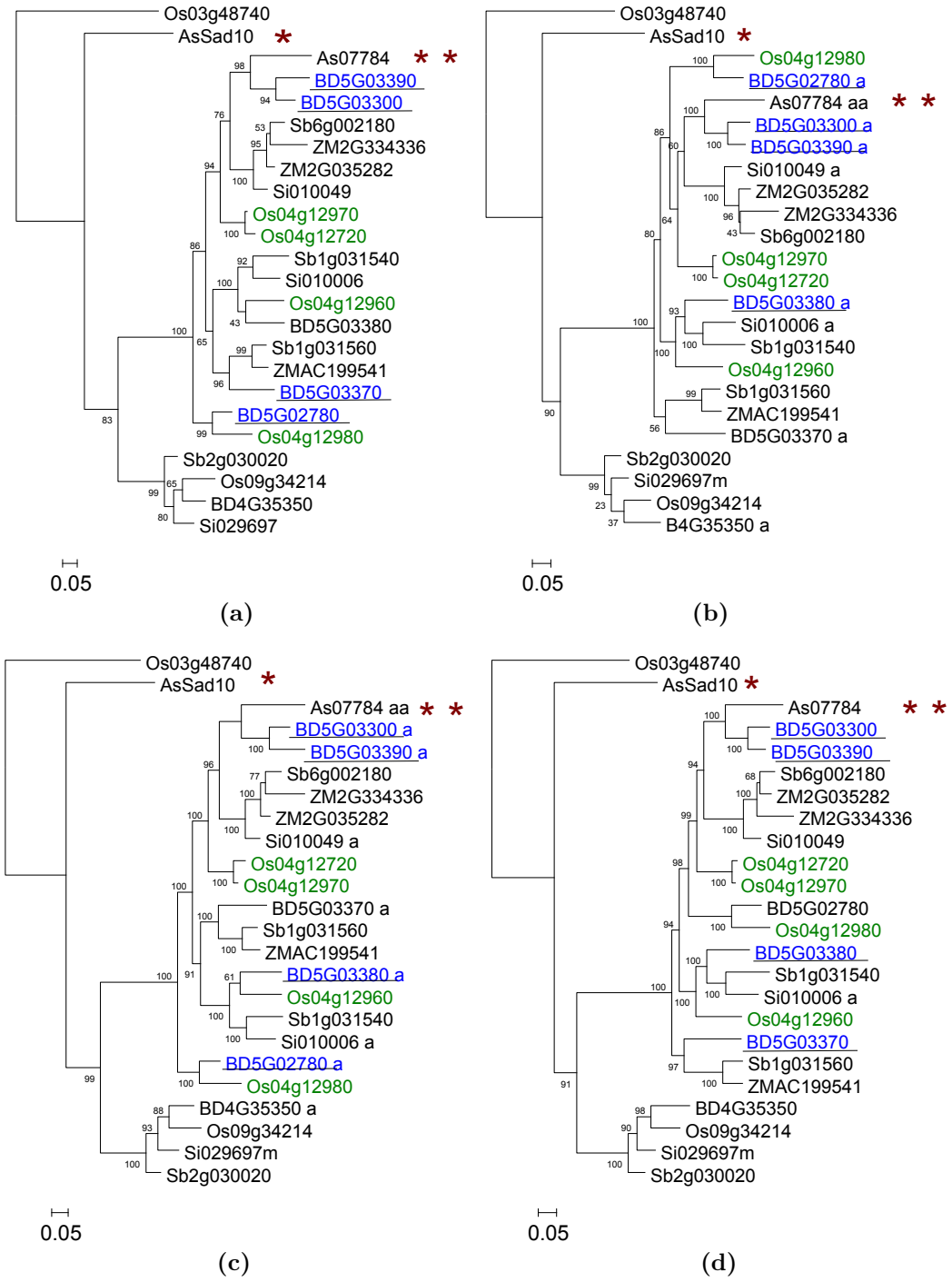


Figure 5.17: Phylogenetic trees of *Sad10* homologues. a) RAxML amino acid tree (lnL = -4727.19, α = 1.11). b) RAxML coding sequence tree (lnL = -9648.87, α = 0.74). c) MrBayes amino acid tree (lnL = -4798.98, α = 1.28). d) MrBayes coding sequence tree (lnL = -10478.99, α = 0.75). *Sad10* (*) and As07784 (* *) are indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 10,000 in RAxML and 75,000,000 in MrBayes) in which the given branching was observed. Tandemly duplicated genes are colour-coded. Original tree files are in the Appendix (Trees 5.26 to 5.29)

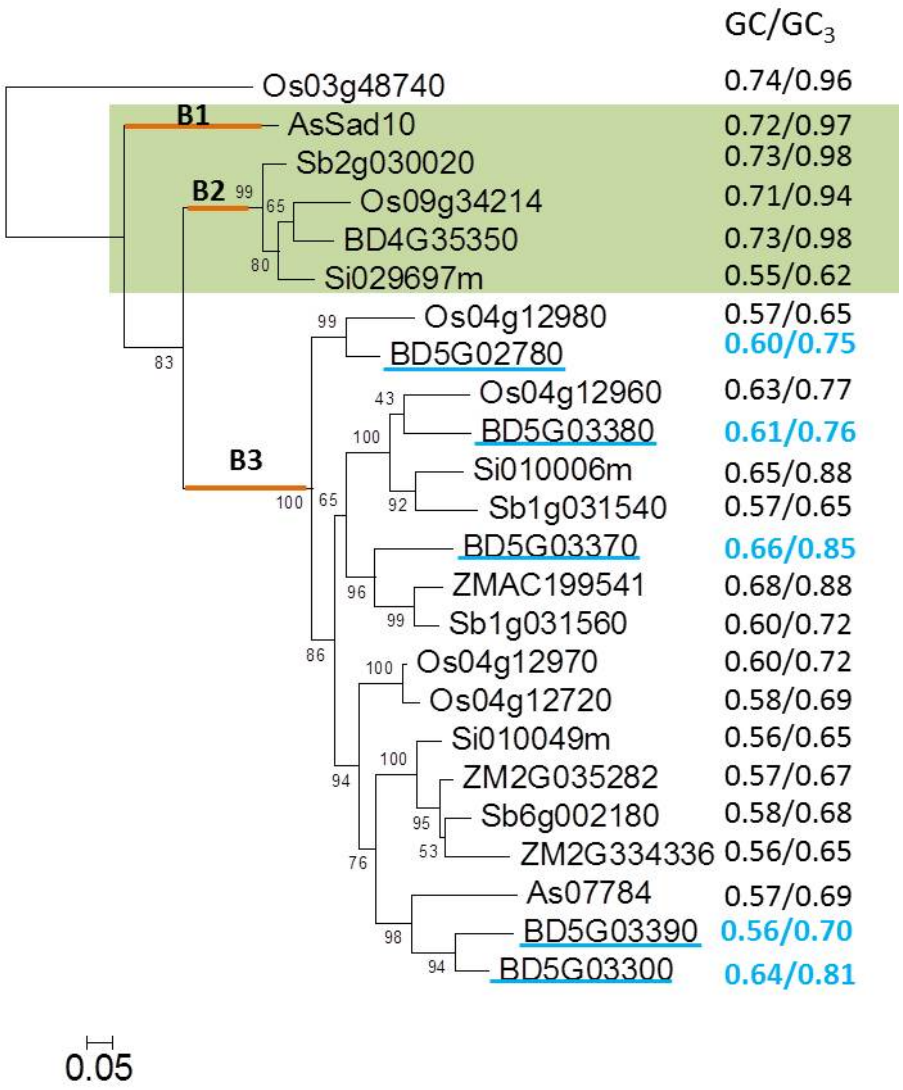


Figure 5.18: The RAxML amino acid *Sad10* homologue tree (Figure 5.17a) with the three branches selected for branch-site tests (B1-B3) labelled. GC/GC_3 contents are listed next to each gene sequence. The tandem array of UGT74s in *B. distachyon* is underlined and their GC contents highlighted in light blue. The clade of genes with elevated GC content is highlighted in green.

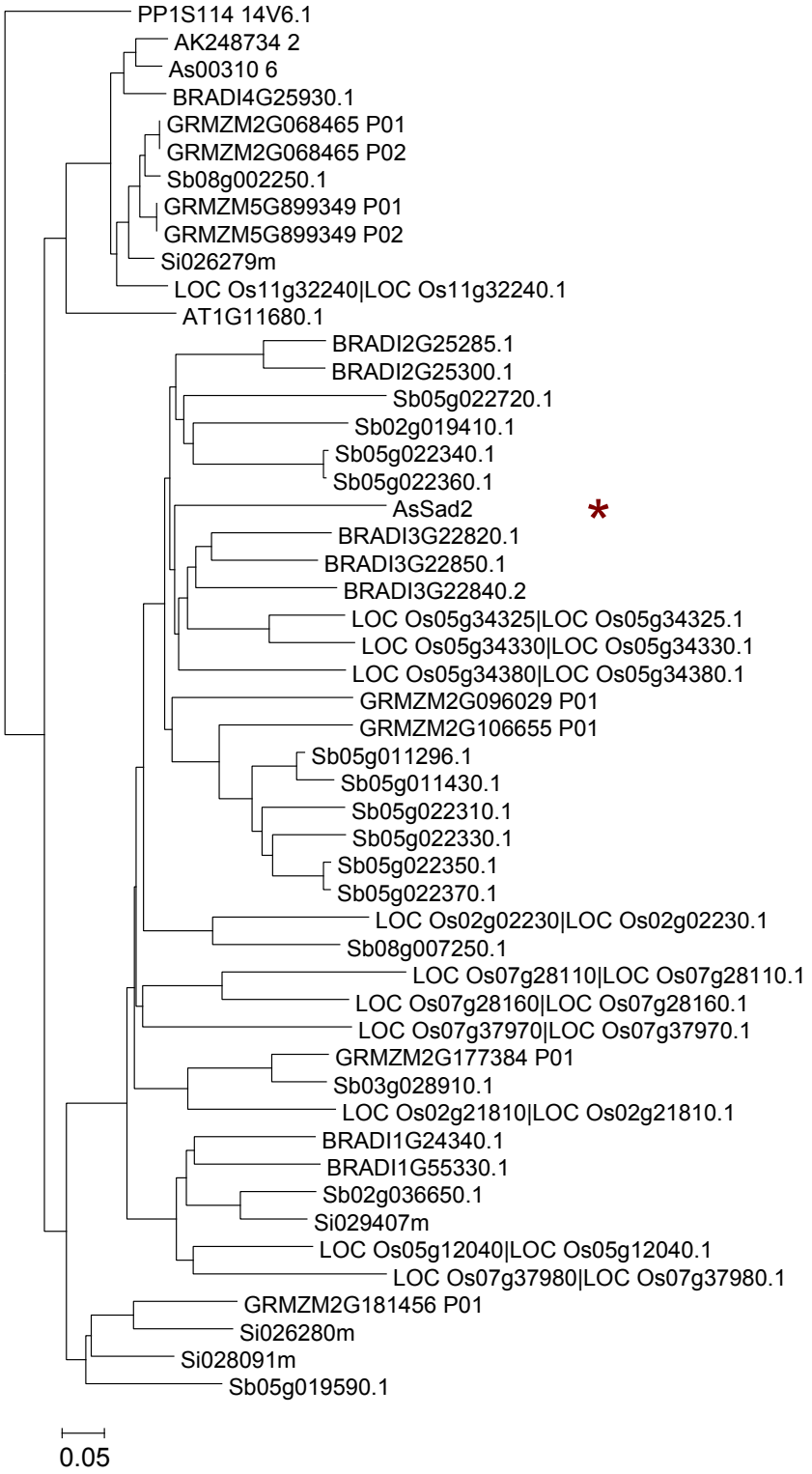


Figure 5.19: The BioNJ tree constructed from the amino acid alignment of CYP51 sequences. *Sad2* is indicated with an asterisk. Original tree file is in the Appendix (Tree 5.30).

5.6.1. Sequence retrieval and annotation of *Sad2* homologues

The amino acid sequences of the closest *A. strigosa Sad2* homologue in *B. distachyon* was identified using BLASTp, and of *O. sativa* genes *CYP51H7* and *CYP51H6* directly from Inagaki et al. (2011). The resulting *Sad2*_closest_hits HMM profile was used to identify 53 hits (E-value <1 x e^{-50}) from the surveyed species: 1 in *A. thaliana*, 2 in *A. strigosa*, 8 in *B. distachyon*, 12 in *O. sativa*, 1 in *P. patens*, 14 in *S. bicolor*, 4 in *S. italica*, 6 in *Z. mays*, and 1 in barley. The HMMsearch did not identify *AtCYP51G2* (AT2G17330) because it is a pseudogene and is not present in the TAIR protein database (Table 5.1). The BioNJ tree resulting from the 51 full-length *CYP51* gene sequences (Appendix Tree 5.30) was compared to the published *CYP51* tree (Inagaki et al., 2011). Following this comparison, all genes in the BioNJ tree were retained for further gene annotation (Figure 5.19). The annotated CYP51 sequences of *O. sativa* and *A. strigosa* were taken directly from Inagaki et al. (2011) and the CYP51 amino acid sequences of *A. thaliana* and *P. patens* were downloaded directly from the relevant Plantensembl databases (Kersey et al., 2012). All other sequences were manually annotated. Four sequences were found to contain in-frame stop codons or to be 3′ truncated, and therefore likely to be non-functional. All but three full-length annotated sequences were found to possess two exons and to contain consistent exon junctions. The details of the gene structures, genomic co-ordinates, gene descriptions, annotated coding sequences and translated amino acid sequences are listed in Appendix Tables 5.13 and 5.14.

5.6.2. Multiple sequence alignment and phylogenetic tree estimation of CYP51 genes

Two sets of alignments were made in order to both maximise the length of the alignment (set 1 with truncated sequences *OsCYP51G4* and *H5* removed, set 2 with both sequences present) while retaining consistency with previous analyses (Inagaki et al., 2011). In addition, five 5′ or 3′ truncated sequences were removed from all alignments, along with highly variable and gapped regions, resulting in 43 and 45 sequences in alignment sets 1 and 2 respectively. The selected evolutionary models for these alignments are shown in Table 5.5.

Phylogenetic trees were estimated using the *P. patens CYP51* gene PP1s114_14V6 (PpCYP51G1) as an outgroup sequence. Both the presence or absence of *OsCYP51G4* and *OsCYP51H5* and the type of alignment affected the topologies of the trees (Figures 5.20a-d and 5.21a-d). A clade of four genes,

Summary of CYP51 phylogenetic trees						
Phylogenetic software	Alignment type	Number of sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	43	410 aa	-15678.04	1.75	JTT + I + Γ + F
RAxML	codon	43	1230 bp	-31060.20	0.92	GTR + Γ
MrBayes	amino acid	43	410 aa	-15772.08	1.74	JTT + I + Γ + F
MrBayes	codon	43	1230 bp	-31126.57	0.92	GTR + Γ
RAxML	3^{rd}_base_ strip	43	820 bp	-15492.88	0.91	GTR + Γ
MrBayes	3^{rd}_base_ strip	43	820 bp	-15564.51	0.91	GTR + Γ
RAxML	amino acid (with OsCYP51G4 and H5)	45	368 aa	-14658.36	1.87	JTT + I + Γ +F
RAxML	codon (with OsCYP51G4 and H5)	45	1104 bp	-28954.59	0.94	GTR + Γ
MrBayes	amino acid (with OsCYP51G4 and H5)	45	368 aa	-14758.23	1.43	JTT + I + Γ + F
MrBayes	codon (with OsCYP51G4 and H5)	45	1104 bp	-28988.63	1.02	GTR + Γ

Table 5.5: Summary of CYP51 phylogenetic analyses.

denoted here as clade 1, was found to be located between the CYP51G and CYP51H genes in the amino acid trees but within the CYP51G group in the codon trees. Inclusion of the *OsCYP51G4* and *OsCYP51H5* sequences in the phylogenetic analysis (Figures 5.20a,b and 5.21a,b) caused the *CYP51G3* group to become distinct from the *CYP51H* genes in all but the MrBayes coding sequence tree. Furthermore, the rice tandem group LOC_Os05g34330, LOC_Os05g34325 and LOC_Os05g34380 were situated in very different parts of the amino acid and coding sequence trees.

In order to resolve this gene tree incongruity, a phylogenetic analysis of the third_base_strip alignment was carried out. Furthermore, the *OsCYP51G4* and *OsCYP51H5* sequences were removed from the analysis. Although the third_base_strip trees (Figure 5.22) showed consistenty between phylogenetic method, they differed both from previously estimated amino acid and coding sequence trees (Figure 5.20 and 5.21). In particular, clade 1 was found within the CYP51G1 clade, similar to the previous coding sequence trees, and the CYP51G3 clade now grouped at the base of the CYP51H clade instead of within it.

To further examine the CYP51 tree topology within the region of *Sad2*, all 25 *CYP51H* genes except for *OsCYP51H5* were included in a *Sad2* homologues phylogenetic tree estimation (Figure 5.23), with LOC_Os05g12040 (OsCYP51G3) included as an outgroup sequence. The topologies of the *AsSad2* clade (Figure 5.23a-d) were consistent with the previous CYP51 trees (Figure 5.20 and 5.21). Sequences closely related to *Sad2* were found to be arranged in tandem on *B. distachyon* chromosomes 2 and 3 and on *S. bicolor* chromosome 5. However, other tandem clusters (e.g. three genes on *O. sativa* chromosome 5) grouped in distinct locations in the *Sad2* homologue trees, indicating potentially older duplication events.

5.6.3.Molecular evolution analysis of *Sad2* homologues

Branch-site tests were performed on four branches of interest in the RAxML amino acid CYP51 phylogenetic tree (Figure 5.24). Each branch was tested under the M1a (foreground $\omega = 1$) and M2a (foreground $\omega > 1$) models. Two positively selected branches, B1 and B4, were identified (Figure 5.24) with the remaining two branches evolving under purifying selection. Detailed branch-site test results are given in Appendix Table 5.15.

The GC content of each gene in the coding sequence alignment was measured (Appendix Table 5.16) (Figure 5.24). The primary sterol biosynthesis CYP51G1s were found to contain a relatively low GC content of approximately 0.55 and

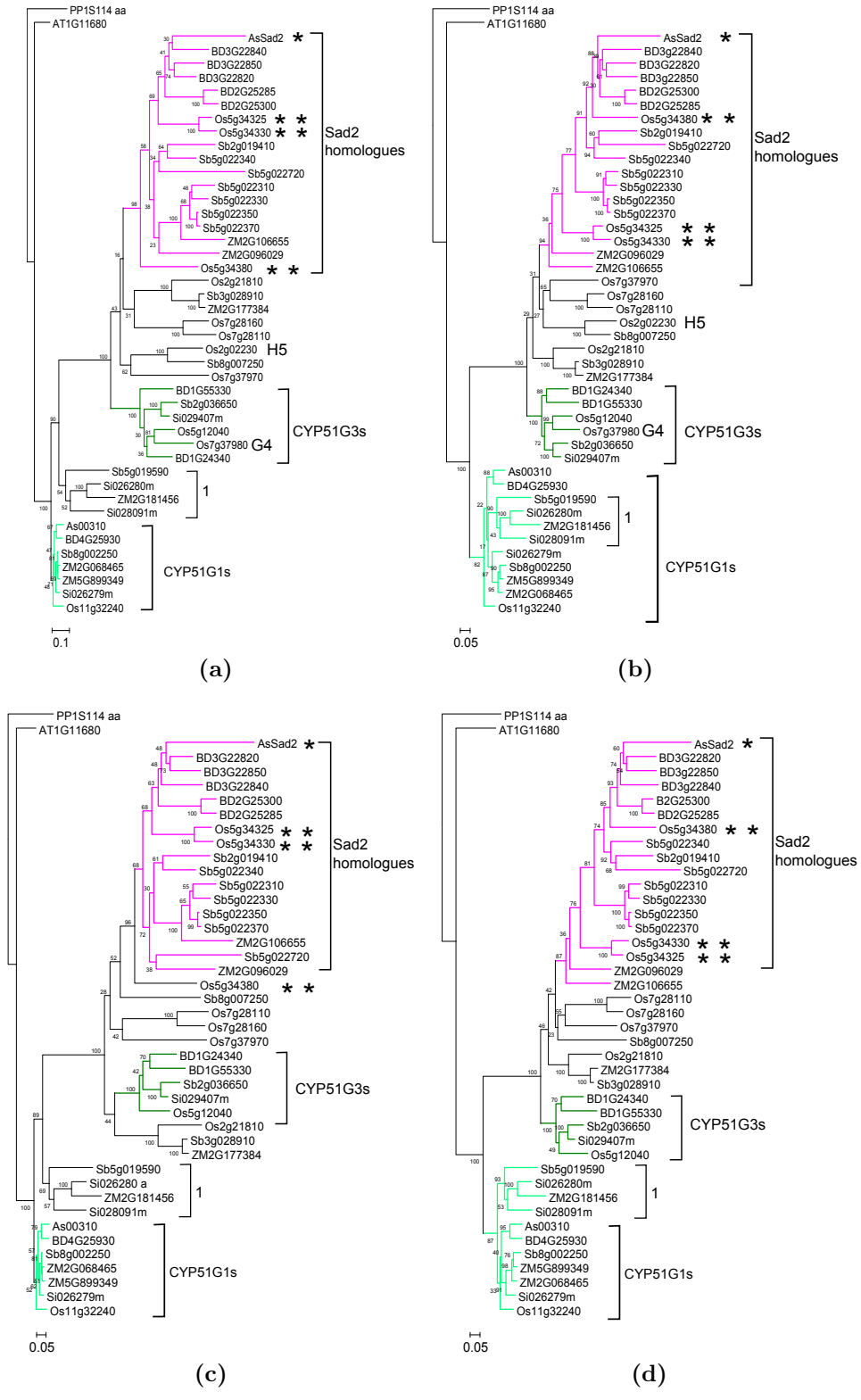


Figure 5.20: The bootstrapped maximum likelihood trees generated in RAxML 7.0.4 using the a) amino acid, b) coding sequence, c) amino acid (without OsCYP51G4 and H5), and d) coding sequence (without OsCYP51G4 and H5) alignments. *CYP51G4* and *H5* are indicated by text. *Sad2* (*), LOC_Os05g34325, 34330 and 34380 (* *) are indicated with asterisks. Bootstrap supports of branches (from 10,000 replicates) are indicated. Original tree files are in the Appendix (Trees 5.31, 5.32, 5.35 and 5.36).

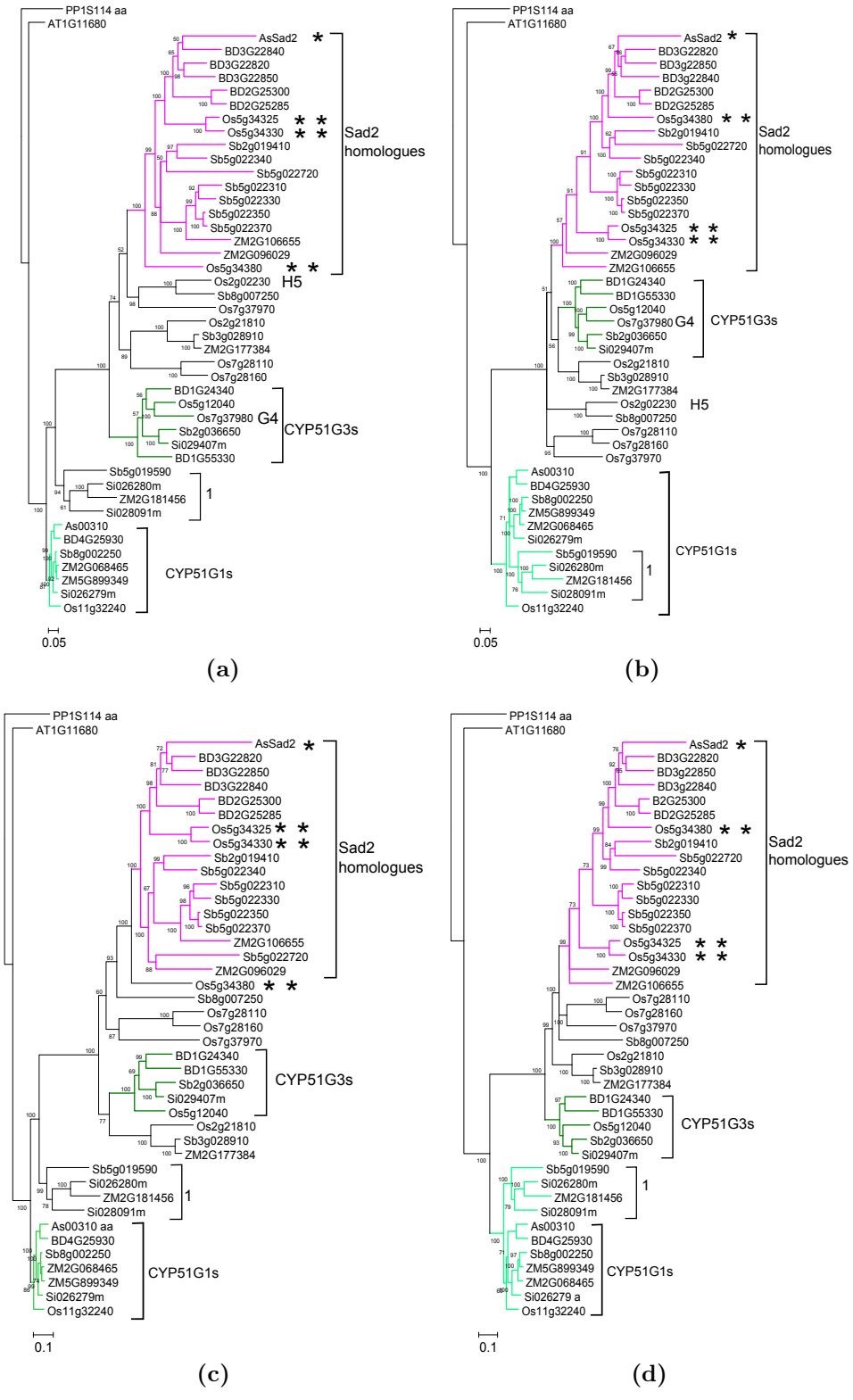


Figure 5.21: The 50 majority consensus trees generated in MrBayes 3.2.1 using the a) amino acid, b) coding sequence, c) amino acid (without OsCYP51G4 and H5), and d) coding sequence (without *OsCYP51G4* and *H5*) alignments. *CYP51G4* and *H5* are indicated by text. *Sad2* (*), LOC_Os05g34325, 34330 and 34380 (* *) are indicated with asterisks. Bootstrap supports of branches (from 75,000,000 MCMC samples) are indicated. Original tree files are in the Appendix (Trees 5.33, 5.34, 5.37 and 5.38)

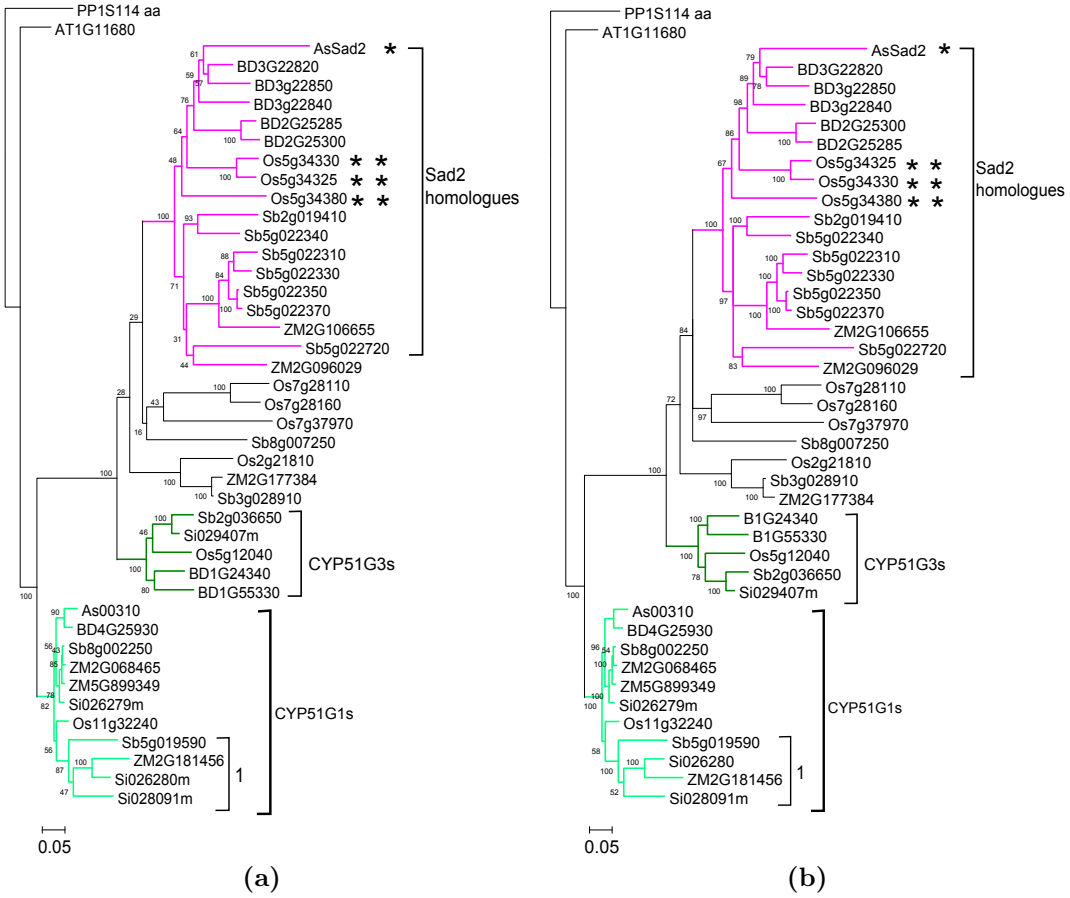


Figure 5.22: CYP51 trees estimated from coding sequence alignment with third base removed and without *OsCYP51G4* and *H5*, in a) RAxML 7.0.4 and b) MrBayes 3.2.1. *Sad2* (*), LOC_Os05g34325, 34330 and 34380 (* *) are indicated with asterisks. Bootstrap supports of branches (from 10,000 and 75,000,000 replicates in RAxML and MrBayes respectively) are indicated. Original tree files are in the Appendix (Trees 5.39 and 5.40).

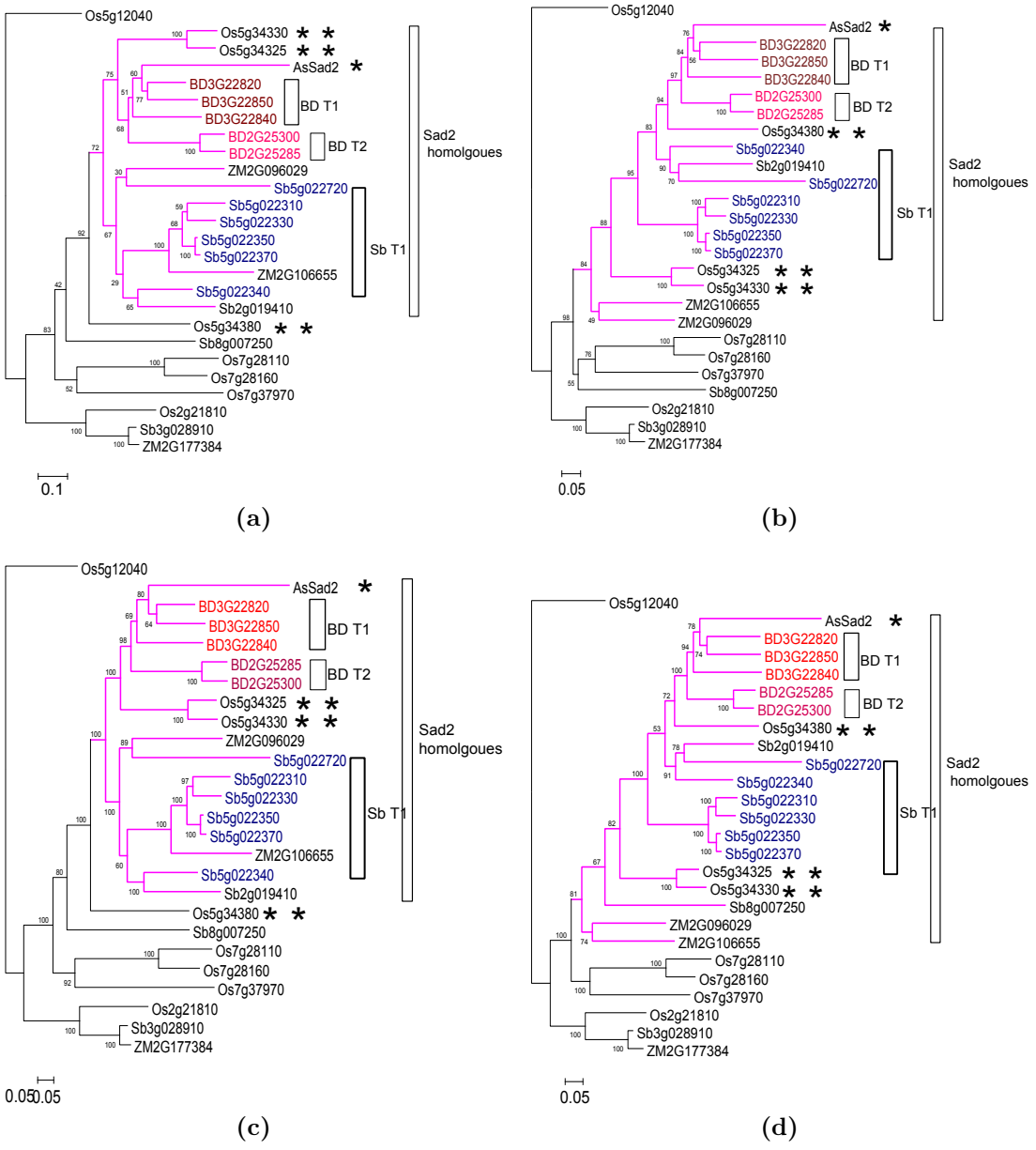


Figure 5.23: Phylogenetic estimation of the *Sad2* homologue tree. a) RAxML 7.0.4 amino acid tree (lnL = -11307.77, α = 1.64). b) RAxML 7.0.4 coding sequence tree (lnL = -20738.59, α = 0.98). c) MrBayes 3.2.1 amino acid tree (lnL = -11382.05, α = 1.64). d) MrBayes 3.2.1 coding sequence tree (lnL = -20780.57, α = 0.96.) *Sad2* (*), LOC_Os05g34325, 34330 and 34380 (* *) are indicated with asterisks. The tandem arrays of *B. distachyon* (BD T1 and BD T2) and that of *S. bicolor* (Sb T1) are indicated. Bootstrap supports of branches (from 10,000 and 75,000,000 replicates in RAxML and MrBayes respectively) are indicated. Original tree files are in the Appendix (Trees 5.41 to 5.44).

a slightly elevated GC_3 content of 0.70, a signature of monocot genes. While the GC content increased in the CYP51G3 clade (average $GC/GC_3 = 0.65/0.80$) and in the subclade of eight genes containing OsCYPH1, it decreased in the *Sad2* homologue clade (Figure 5.24). *Sad2* itself exhibited a GC/GC_3 content considerably lower than the monocot CYP51G1s and very similar to that of the *A. thaliana* CYP51G1 gene.

5.7.Results of phylogenetic analysis of *Sad1*

The evolution of plant triterpene synthases has been studied in detail by Xue et al. (2012). The authors showed that monocot triterpene synthases, including *Sad1*, derived via duplication-neofunctionalization of an ancestral cycloartenol synthase while dicot triterpene synthases derived from lanosterol synthases. Furthermore, tandem duplications played an important role in the expansion of the *OSC* gene family in both dicots and monocots. Xue et al. (2012) also showed that relaxed selection was frequently observed on one of the two branches following an *OSC* duplication event.

Sad1 encodes the 'signature enzyme' for avenacin biosynthesis (Figure 1.2a) and is the key enzyme that diverts the production of sterol primary metabolism to triterpene biosynthesis (Chu et al., 2011). Because SAD1 is the first committed enzyme of the avenacin biosynthetic pathway, its evolution led to the birth of avenacin biosynthesis and may have stimulated subsequent gene recruitment events of the avenacin pathway 'tailoring enzymes'. Here, the previous analyses of *Sad1* are extended, to shed further light on the evolution of the OSC gene family.

5.7.1.Sequence retrieval and annotation of *Sad1* homologues

The amino acid sequences of *A. strigosa* S75 *Sad1* and its closely related homologues, LOC_Os06g28820, BRADI1G42000, and Sb10g029175 were used to build a *Sad1*_closest_hits HMM profile. The subsequent HMMsearch identified a total of 126 hits: 16 in *A. thaliana*, 8 in *B. distachyon*, 18 in *Z. mays*, 19 in *O. sativa*, 1 in *P. patens*, 1 in wheat, 33 in *S. bicolor*, 25 in *S. italica*, 3 in *A. strigosa* and 2 in barley. The search did not identify LOC_Os08g12740 (OsOSC12), because it is not represented in the MSU6 protein database. An additional tblastx search of the *AsSad1* protein sequence against the corresponding genomic sequence databases identified two further loci in *S. bicolor*, three in *S. italica*, one in *B. distachyon*, and one (OsOSC12) in rice, but none in *Z. mays*. The newly

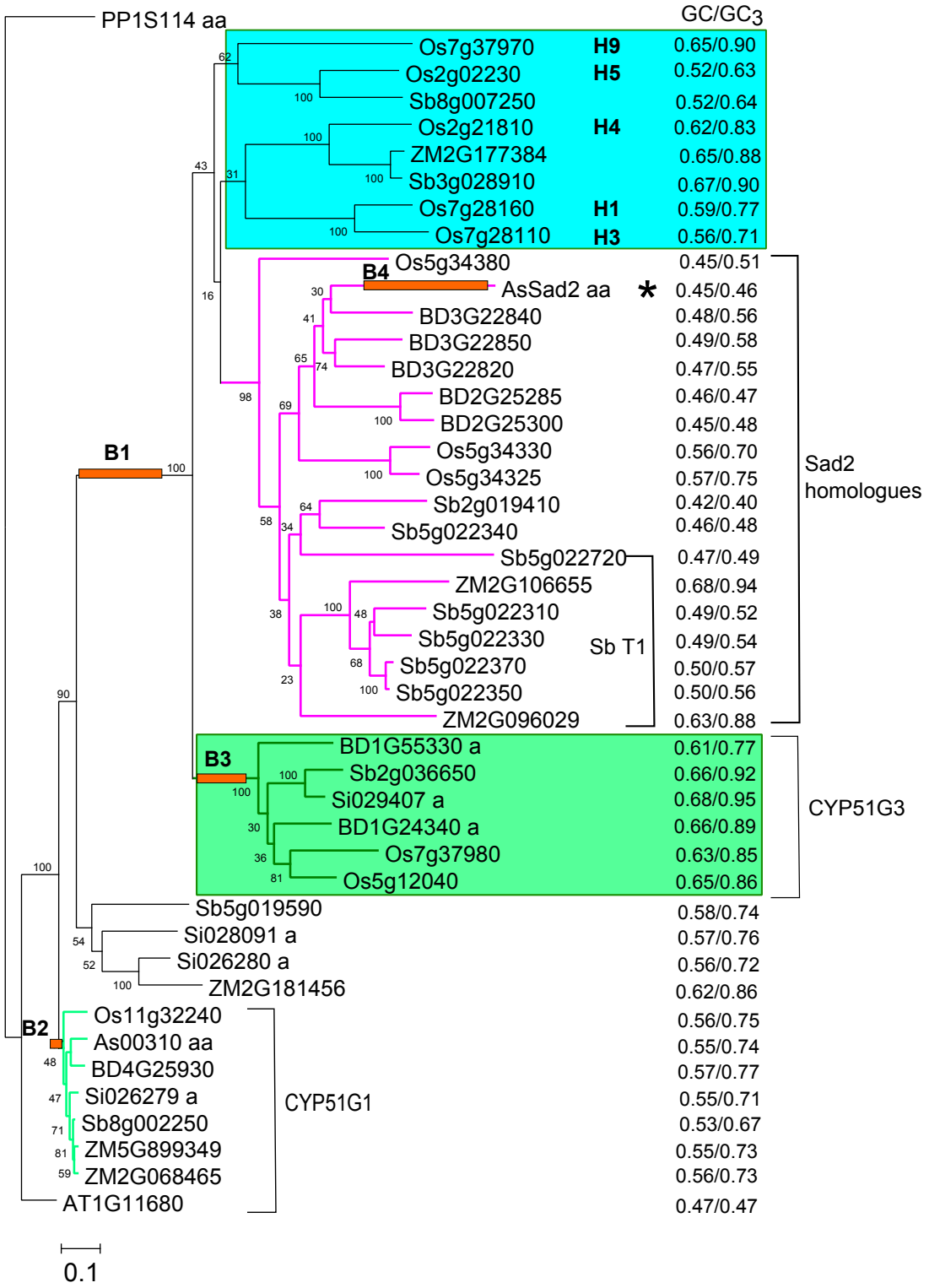


Figure 5.24: The RAxML amino acid *Sad2* homologue tree (Figure 5.20a) with the four branches selected for branch-site tests (B1-B4) labelled. GC/GC_3 contents are listed next to each gene sequence. *Sad2* is indicated with an asterisk. The tandem array in *S. bicolor* (Sb T1) is indicated. Clades with elevated GC content are highlighted.

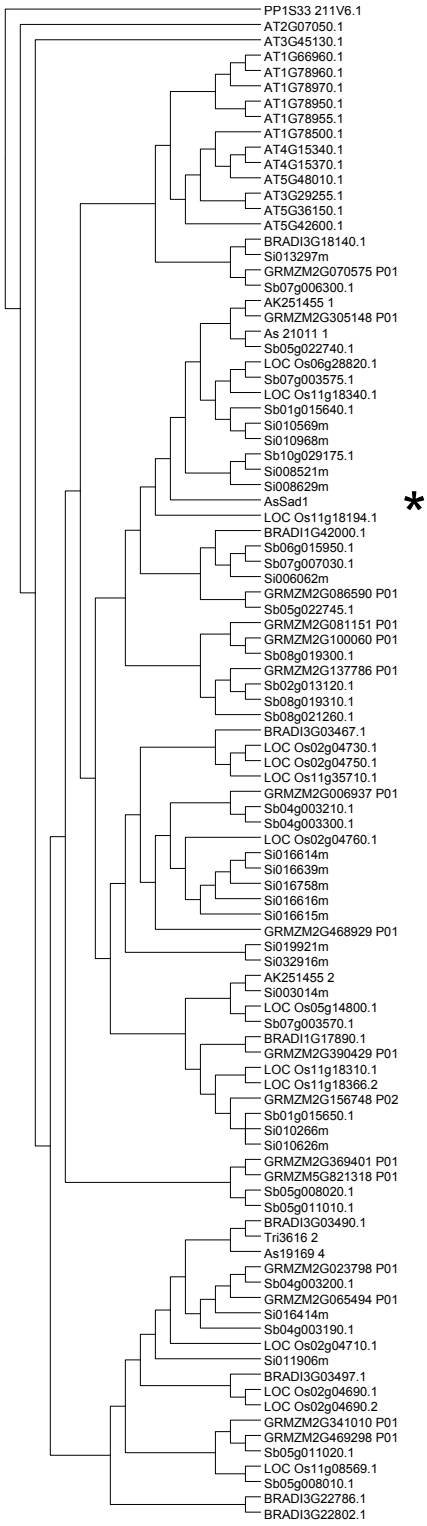


Figure 5.25: The BioNJ tree estimated from the amino acid alignment of 99 OSCs with an E-value <1 x e^{-50}. *Sad1* is indicated with an asterisk. Original tree file is in the Appendix (Tree 5.45).

discovered hits were all included in the subsequent phylogenetic analysis but not in BioNJ tree construction. A BioNJ tree of 99 hits with an E-value <1 x e^{-50} (Figure 5.25) was compared to two previously published oxidosqualene cyclase (OSC) phylogenetic trees (Inagaki et al., 2011; Xue et al., 2012). All hits were verified as *OSCs* and were thus annotated for later phylogenetic analysis.
Gene annotation of the monocot *OSCs* revealed that most genes share an 18-exon structure, with the exception of five full-length genes with 16 or 17 exons. In the *Z. mays, S. bicolor*, and *S. italica* genomes, several OSCs possess very long introns (>10,000 bp) and are indicated as two loci in their respective genome databases. However, the majority of tblastx-identified OSCs are likely to be pseudogenes due to truncations or frame-shift mutations. Most *A. thaliana OSCs* were found to consist of either 14 or 17 exons, one full-length gene (AT5G36150) possesses only 13 exons, and another (AT3G29255) in likely to be truncated. The gene structures, genomic coordinates, and sequences spanning the exon junctions of the annotated genes are listed in Appendix Table 5.16. The annotated coding sequences and translated amino acid sequences are listed in Appendix Table 5.17.

5.7.2. Multiple sequence alignment and phylogenetic tree estimation of OSC sequences

Following removal of truncated OSC sequences and those possessing highly variable regions, a total of 70 full-length amino acid sequences were aligned. The corresponding coding sequence alignment was generated in Pal2Nal (Suyama et al., 2006) using the amino acid alignment as a reference. The selected evolutionary models for these alignments are shown in Table 5.6.

Summary of OSC phylogenetic trees						
Phylogenetic software	Alignment type	Number of sequences	Alignment length	(Mean) lnL	(Mean) α	Model
RAxML	amino acid	70	555 aa	-27451.61	1.64	JTT + I + Γ + F
RAxML	codon	70	1665 bp	-57637.96	0.82	GTR + Γ
MrBayes	amino acid	70	555 aa	-27603.90	1.12	JTT + I + Γ + F
MrBayes	codon	70	1665 bp	-57709.12	0.81	GTR + Γ

Table 5.6: Summary of OSC phylogenetic analyses

The four phylogenetic trees exhibit consistent topologies, differing only in the the groupings of the clades containing LOC_Os11g08569 and LOC_Os02g04690 (Figures 5.26a,b and 5.27a,b). When compared to the reference OSC tree (Xue et al., 2012), the group of close *Sad1* homologues (pentacyclic triterpene synthase-like) and the cycloartenol synthase (*CAS*) clade were identified (Figures 5.26 and 5.27). The *Sad1* homologue clade does not strictly follow the

known species phylogeny, as *AsSad1* groups more closely with *Panicoideae* OSC sequences than to those of rice and *B. distachyon*. Furthermore, the OSC sequence BRADI1G42000 was found to be grouped within a clade somewhat distant from the *Sad1* homologue clade, leaving the question of whether *B. distachyon* possesses a *Sad1* orthologue unanswered. In contrast to the large pentacyclic triterpene synthase-like group (containing 23 sequences), all of which possess relatively long branches, the CAS clade contains only nine sequences and all branches are short (Figure 5.26 and 5.27). Furthermore, the CAS clade was found to follow the species phylogeny, supporting its role as an ancient primary metabolic gene. No dicot genes were identified in the CAS clade, indicating that it is a monocot-specific gene and reconfirming the previous hypothesis that the ancient *CAS* gene was lost in dicots (Xue et al., 2012).
A phylogenetic estimation of a smaller set of *Sad1* homologues was carried out, rooted with the outgroup sequence AT2G07050. All four *Sad1* homologues trees share largely consistent topologies (Figure 5.28), differing only in the arrangement of the clades containing LOC_Os02g04690, LOC_OS08g12740 and LOC_Os11g08569, leaving their true location somewhat uncertain.

5.7.3.Selection tests of *Sad1* homologues

Branch-site tests were performed on eight branches of interest in the *Sad1* homologue tree (Figure 5.29). These branches represented key duplication events leading to the evolution of either cycloartenol synthases or β-amyrin synthases (Appendix Table 5.18). Branches B1, B2, B4, B5, and B6 were found to have evolved under positive selection (Figure 5.29), while the other tested branches were under purifying selection.
The GC contents of each gene in the *Sad1* homologues coding sequence alignment were measured (Appendix Table 4.19) and found to range from 0.44 to 0.53 (Figure 5.29), with no clear difference between primary and secondary metabolic OSCs. More interestingly, the GC content of *AsCAS* (0.48/0.47) was found to be slightly higher than that of *AsSad1* (0.41/0.44). The GC contents of the monocot OSCs were unusually low and very similar to the GC content of the *A. thaliana* lanosterol synthase. The clade possessing the highest GC content (marked in green in Figure 5.29), exhibited a GC_3 content approximately 10% higher than all pentacyclic triterpene synthase-like sequences.

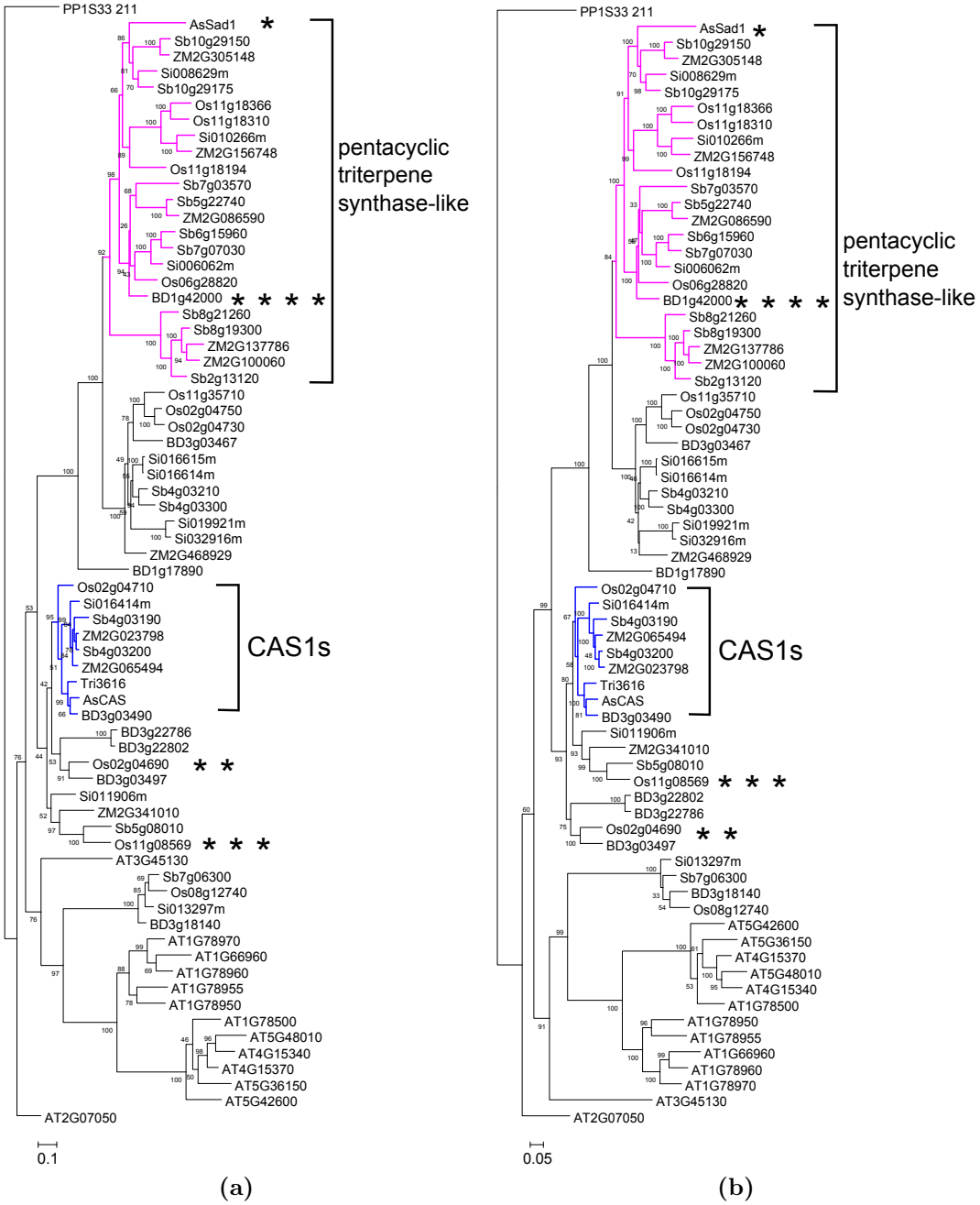


Figure 5.26: The bootstrapped maximum likelihood trees generated in RAxML 7.0.4 using the a) amino acid and b) coding sequence alignments of 70 OSCs. *Sad1* (*), LOC_Os02g04690 (* *), LOC_Os11g08569 (* * *) and BRADI1G42000 (* * * *) are indicated with asterisks. The numbers indicate the percentage of bootstrap replicates (out of 10,000) in which the given branching was observed. Original tree files are in the Appendix (Trees 5.46 and 5.47).

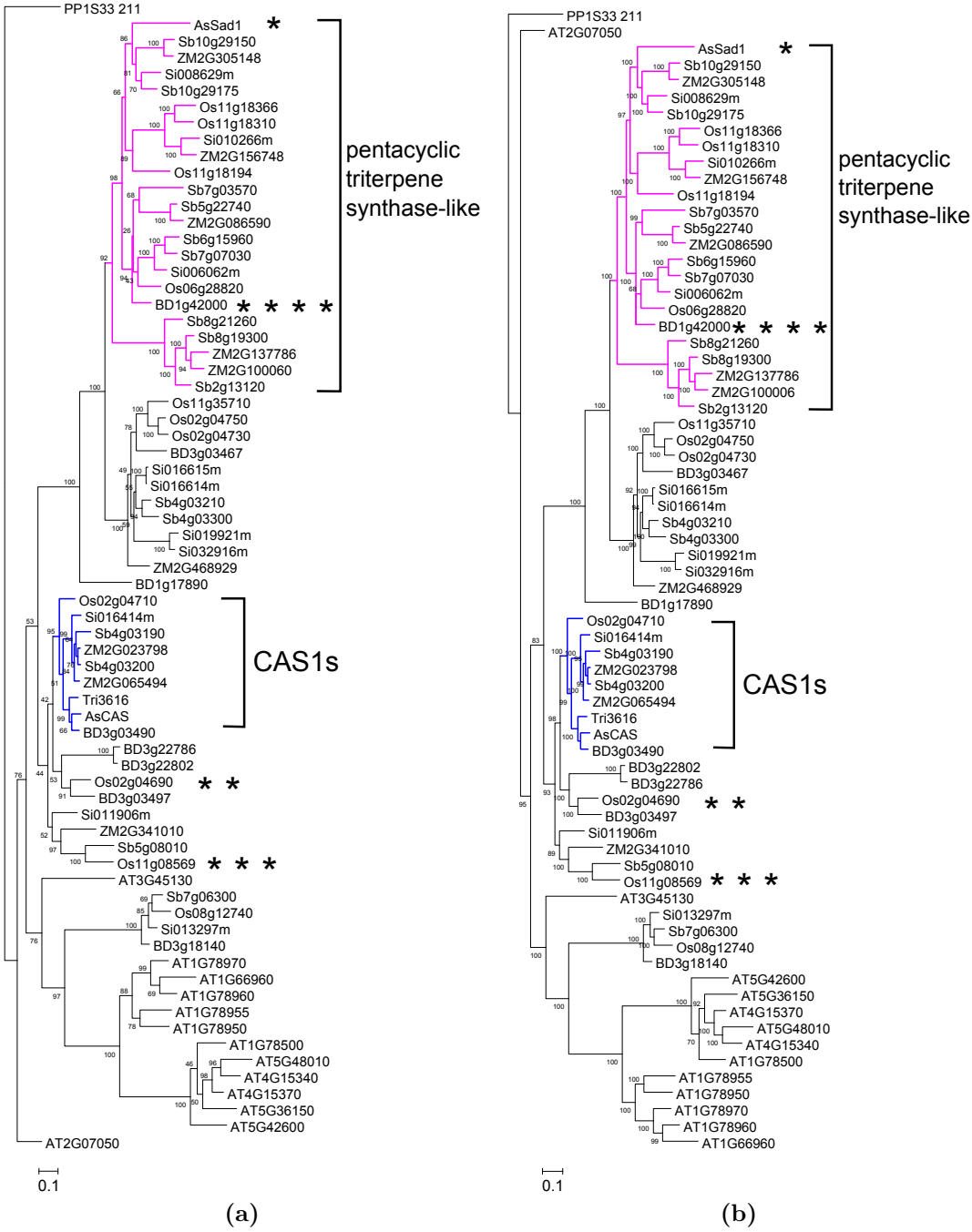


Figure 5.27: The 50 majority consensus trees generated in MrBayes 3.2.1 using the a) amino acid and b) coding sequence alignments of 70 OSCs. *Sad1* (*), LOC_Os02g04690 (* *), LOC_Os11g08569 (* * *) and BRADI1G42000 (* * * *) are indicated with asterisks. The bootstrap supports of branching patterns (from 75,000,000 MCMC samples) are indicated. Original tree files are in the Appendix (Trees 5.48 and 5.49).

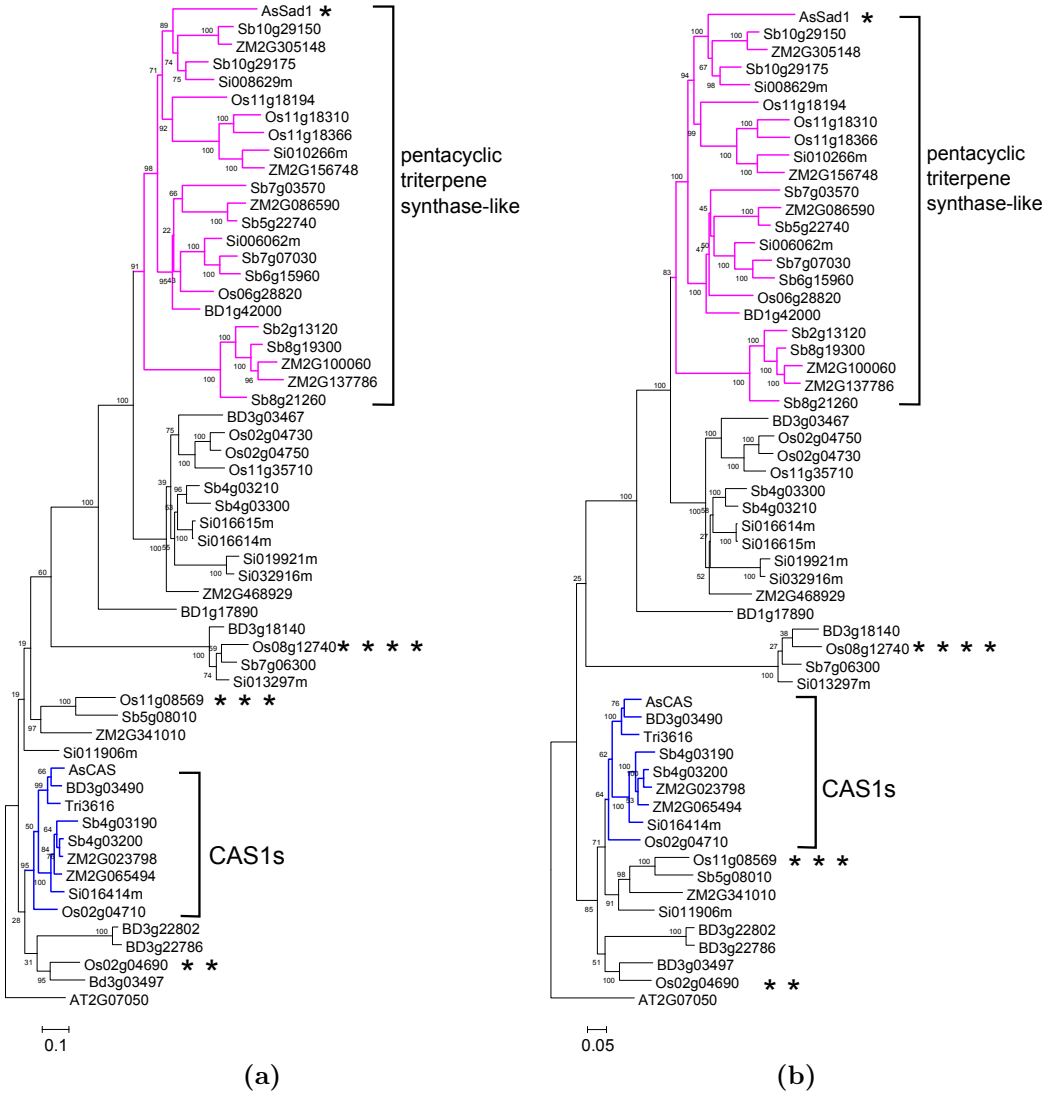


Figure 5.28: Phylogenetic tree estimation of the *Sad1* homologue trees. a) RAxML amino acid tree (lnL= -20880.49, α= 1.50), b) RAxML coding sequence tree (lnL= -44077.63, α= 0.81), c) MrBayes amino acid tree (lnL= -21020.32, α= 1.07) and d) MrBayes coding sequence tree (lnL= -44145.80, α= 0.81). Bootstraps supports of branches (from 10,000 and 75,000,000 replicates in RAxML and MrBayes respectively) are indicated. *Sad1* (*), LOC_Os02g04690 (* *), LOC_Os11g08569 (* * *) and LOC_Os08g12740 (* * * *) are indicated with asterisks. Original tree files are in the Appendix (Trees 5.50 to 5.53).

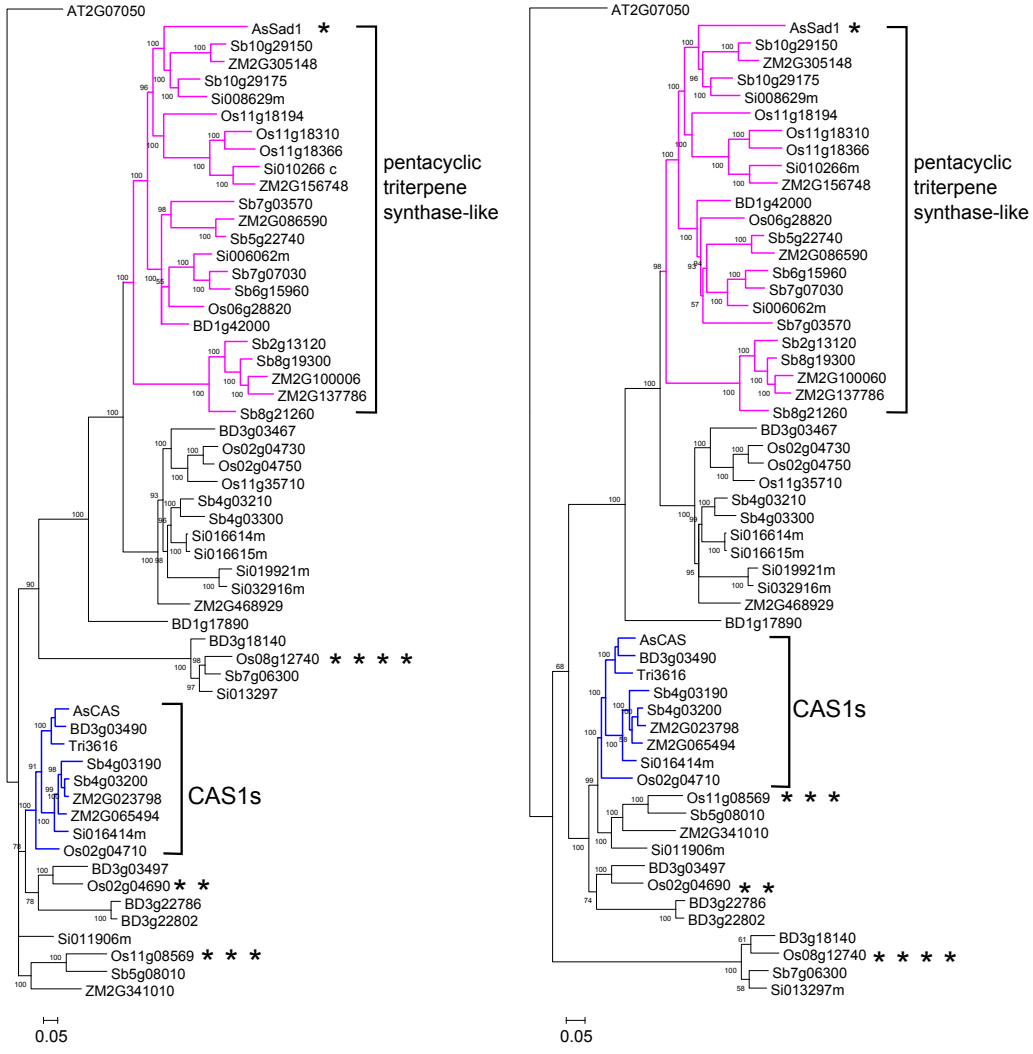

AT2G07050
AsSad1 *
Sb10g29150
ZM2G305148
Sb10g29175
Si008629m
Os11g18194
Os11g18310
Os11g18366
Si010266 c
ZM2G156748
Sb7g03570
ZM2G086590
Sb5g22740
Si006062m
Sb7g07030
Sb6g15960
Os06g28820
BD1g42000
Sb2g13120
Sb8g19300
ZM2G100006
ZM2G137786
Sb8g21260
pentacyclic triterpene synthase-like
BD3g03467
Os02g04730
Os02g04750
Os11g35710
Sb4g03210
Sb4g03300
Si016614m
Si016615m
Si019921m
Si032916m
ZM2G468929
BD1g17890
BD3g18140
Os08g12740 * * * *
Sb7g06300
Si013297
AsCAS
BD3g03490
Tri3616
Sb4g03190
Sb4g03200
ZM2G023798
ZM2G065494
Si016414m
Os02g04710
CAS1s
BD3g03497
Os02g04690 * *
BD3g22786
BD3g22802
Si011906m
Os11g08569 * * *
Sb5g08010
ZM2G341010
0.05
(c)
AT2G07050
AsSad1 *
Sb10g29150
ZM2G305148
Sb10g29175
Si008629m
Os11g18194
Os11g18310
Os11g18366
Si010266m
ZM2G156748
BD1g42000
Os06g28820
Sb5g22740
ZM2G086590
Sb6g15960
Sb7g07030
Si006062m
Sb7g03570
Sb2g13120
Sb8g19300
ZM2G100060
ZM2G137786
Sb8g21260
pentacyclic triterpene synthase-like
BD3g03467
Os02g04730
Os02g04750
Os11g35710
Sb4g03210
Sb4g03300
Si016614m
Si016615m
Si019921m
Si032916m
ZM2G468929
BD1g17890
AsCAS
BD3g03490
Tri3616
Sb4g03190
Sb4g03200
ZM2G023798
ZM2G065494
Si016414m
Os02g04710
CAS1s
Os11g08569 * * *
Sb5g08010
ZM2G341010
Si011906m
BD3g03497
Os02g04690 * *
BD3g22786
BD3g22802
BD3g18140
Os08g12740 * * * *
Sb7g06300
Si013297m
0.05
(d)

Figure 5.28

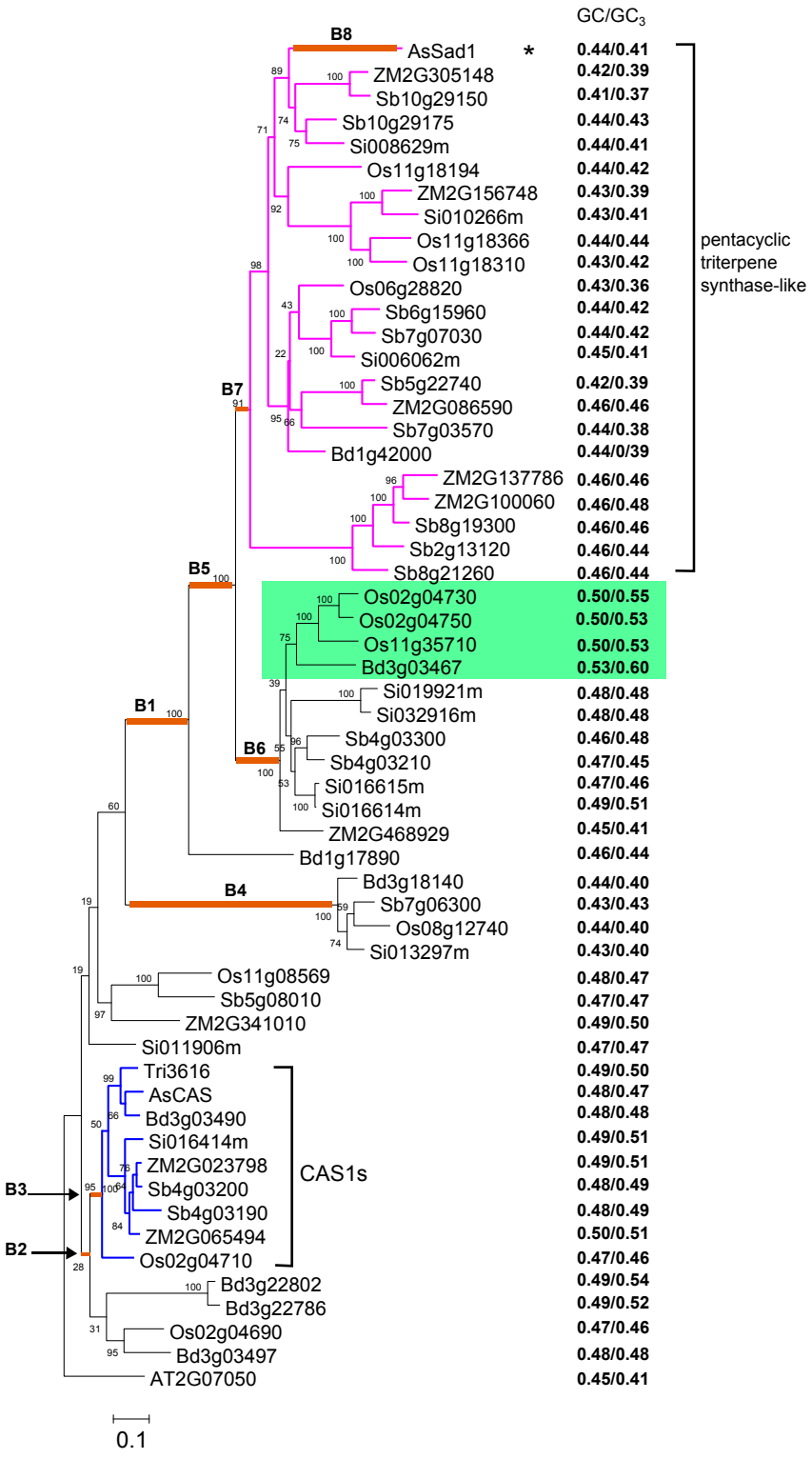


Figure 5.29: The RAxML amino acid *Sad1* homologue tree (Figure 5.28) with the eight branches selected for branch-site tests (B1-B8) labelled. *Sad1* is indicated with an asterisk. GC/GC_3 contents are listed next to each gene sequence. The clade of highest GC content is higlighted in green.

5.8.Discussion

Genome mining of *Sad* gene homologues in six annotated higher plant genomes, the *P. patens* genome, and three fl-cDNA databases was carried out in order to investigate the evolution of individual components of the avenacin biosynthetic gene cluster. This led to identification of 70 OSCs, 43 CYP51s, 86 SCPLs Clade1A, 134 Class I OMTs, 105 Group L UGTs amongst the surveyed species. Through careful gene annotation, construction of multiple sequence alignments and phylogenetic trees, key features of the *Sad* gene evolutionary histories were identified and the phylogenetic relationships of the *Sad* genes and their homologues were elucidated.

5.8.1.The improved analytical pipeline

Identification of *Sad1* and *Sad2* homologues in previous studies using BLASTn and tBLASTx (Gish, 1994) was relatively straightforward due to the small gene family size (Inagaki et al., 2011; Xue et al., 2012). However, even in the case of *Sad1* an iterative BLAST search, where hits from a previous round of analysis were used to identify further hits, was necessary to find all gene family members (Inagaki et al., 2011). Such an approach would be labour intensive for large, diverse (both in terms of sequence and function) gene families such as *Sad7, 9*, and *10*. Furthermore, closely related but truncated or misannotated sequences resulting in a short sequence alignment to a query sequence might result in a low BLAST E-value, leading to omission of the sequences from further analysis.
Replacing BLAST with HMMER search (Eddy, 2009), using HMM profiles built from the closest homologues of each *Sad* gene, greatly increased the sensitivity and specificity of the searches, and an iterative searching strategy proved unnecessary. For example, the HMMER search of *Sad7* homologues retrieved 254 hits from the ten surveyed databases, compared to 36 hits identified using BLASTn (Gish, 1994) with the same E-value threshold. However, the HMMER searches carried out here relied heavily on the quality of genome annotation and it was not possible to identify either pseudogenes lacking a translated protein sequence or misannotated loci. While additional searches for such sequences were not made here for the cases of *Sad7, 9*, and *10*, such an analysis could be carried out in the future, as the quality of gene annotations in sequenced genomes improves.
Restricting phylogenetic tree construction to full-length coding sequences may not have fully captured the evolutionary history of the *Sad* gene homologues, because pseudogenization or loss of alternative exons are common fates of gene duplicates (Innan and Kondrashov, 2010; Lynch and Force, 2000). However,

inclusion of truncated sequences would reduce both the length and quality of multiple sequence alignment and increase both the analytical time and the uncertainty of tree topology estimation. For the majority of analyses, truncated sequences were excluded in order to generate reliable phylogenies upon which inference of evolutionary events could be made. The only exception from this strategy was the analysis of *CYP51s*, where a limited number of truncated sequences were kept in order to maintain consistency with the previous phylogenetic study (Inagaki et al., 2011).
Construction of a BIONJ tree from the HMMER search output enabled immediate identification of the clade(s) of relevant homologues, through comparison to reference trees, and facilitated the determination of a suitable outgroup sequence for subsequent phylogenetic tree estimation. However, for less well characterised gene families, a tool such as OrthoMCL (Li et al., 2003) might be more appropriate.

5.8.2.Most *Sad* genes emerged after the monocot/dicot split

Analysis of the *Sad1, 2, 7, 9,* and *10* phylogenetic trees revealed that none of the *Sad* genes possessed closely related homologues in dicots, suggesting their emergence occurred after the monocot-dicot divergence. This view is consistent with Xue et al. (2012), who showed that *Sad1* derived from the monocot ancestral CAS, and a previous study indicating that *Sad7* was monocot-specific (Mugford and Osbourn, 2010). This phylogenetic analysis has shown that the ancestral forms of *Sad1, 2* and *7* are relatively ancient, and that their emergence took place before the grass radiation. However, due to the extensive expansions of the OMT and UGT gene families, the origins of *Sad9* and *Sad10* are not clear. In the *Sad9* homologous clade, no true orthologues of *Sad9* in *S. bicolor, S. italica*, and *Z. mays* were found. Therefore, *Sad9* may have emerged only within the subfamilies Bambusoideae, Ehrhartoideae and Pooideae (the BEP clade). *Sad10* was found to be located at the base of the *Sad10* homologues clade and shared no closely related homologues (except in the 3rd_base_strip trees) with other species investigated.

5.8.3.Tandem duplication played an important role in the specialization of *Sad* genes

Several OSC, CYP51, Clade1A SCPL, Class I OMT and Group L UGT gene family members were found to be arranged in tandem arrays within the genomes of *B. distachyon*, *O.sativa* and *S. bicolor*. However, the locations of the *Z. mays*

homologues were more scattered across the genome, likely due to frequent gene movements and genome rearrangements in that species (Zhang et al., 2011). Many tandem duplicates were grouped into distinct but neighbouring clades of the phylogenetic trees, suggesting structural and functional divergence. While *Sad1, Sad7* and *Sad10* homologues were in each case in tandem before the grass radiation, BEP clade-specific or lineage-specific tandem duplications were seen to have occurred in the *Sad2* tree. The prevalence of tandemly arranged triterpene biosynthesis genes is consistent with the observation that tandem duplicates experience accelerated divergence and are selectively retained due to advantageous new functionality (Hanada et al., 2008; Rodgers-Melnick et al., 2012; Wang, 2013). It could be speculated that terpenoid diversity is generally important for plant adaptation, therefore driving the divergence of terpene biosynthetic genes through tandem duplication.

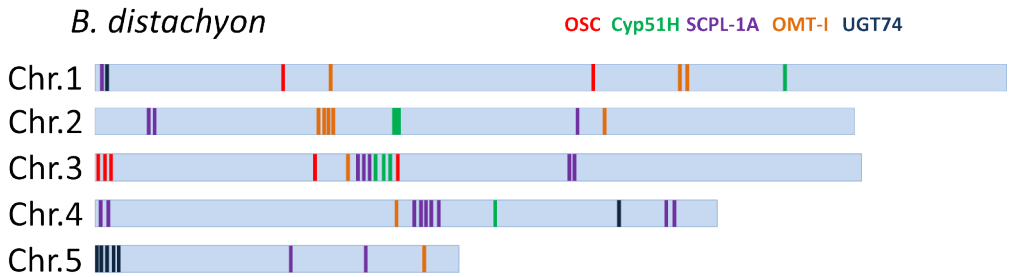
5.8.4. No avenacin-related gene clusters were found in other sequenced monocot genomes

Although closely related homologues of *Sad1, Sad2*, and *Sad7* avenacin biosynthesis genes are present in *B. distachyon, O. sativa*, and *S. bicolor*, neither homologous nor convergent avenacin gene clusters were identified. Nonetheless, OSCs, CYP51s, SCPLs Clade1A and class I OMTS were found to be enriched on *B. distachyon* chromosome 3 and *O. sativa* chromosome 11 (Figure 5.30), suggesting that these triterpene biosynthetic genes may be predisposed to physical linakge.

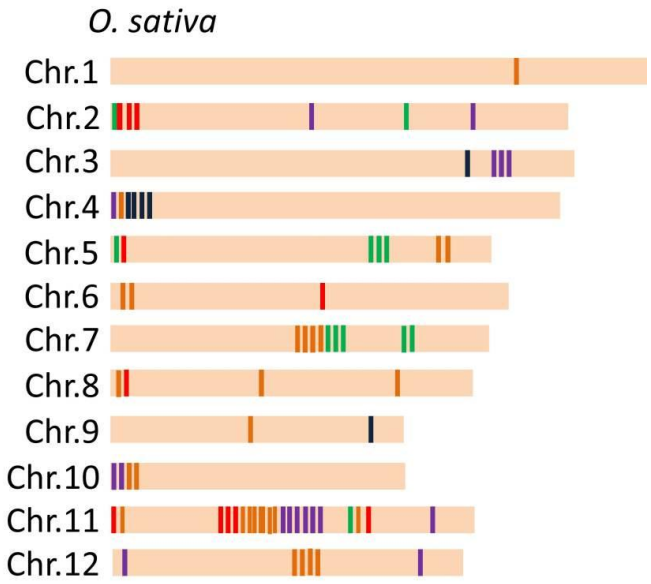
5.8.5. GC_3 content may be affected by elevated levels of synonymous substitution

Dramatic differences in GC contents were observed within the homologues of *Sad2* and *Sad7*. *Sad9* and *Sad10* homologues exhibited moderate GC_3 variation, with only small differences of GC contents amongst clades. *Sad1* showed the smallest GC/GC_3 variation of the five gene families. Interestingly, a gradual decrease in GC/GC_3 content was exhibited on the evolutionary path from primary metabolic *CAS* to secondary metabolic *Sad1*.

Sad1 and *Sad2* contain exceptionally low GC/GC_3 content amongst monocot genes, suggesting either that they are located within GC-poor genomic regions or that selection for low GC content has occurred during their evolution. In contrast, the GC_3 contents of *Sad7* and its close homologues were almost double those of their tandem paralogues, suggesting that *Sad7*-like genes may have



(a)



(b)

Figure 5.30: The distribution of triterpene biosynthetic genes in a) *B. distachyon* and b) *O. sativa* genomes. OSCs are highlighted in red, CYP51Hs in green, SCPL1s in purple, Class I OMTs in orange, and UGT74s in blue.

undergone selection for high GC_3 content, through synonymous mutations due to selection on codon usage (Tatarinova et al., 2010), or extensive GC-biased gene conversion (gBCG) recombination. GC richness has been found to be correlated with higher *dN*, *dS* rates and GC_3 content (Jiang et al., 2013; Serres-Giardi et al., 2012). Genes possessing high GC content may be prone to DNA methylation and thus differential expression, which is important for gene retention and divergence (Jiang et al., 2013; Lukens et al., 2006). It could be speculated that tandem arrays of Class I OMTs and Group L UGTs achieved gene family expansion and functional divergence through an increase in GC content and accelerated synonymous changes (Jiang et al., 2013; Wang, 2013).

5.8.6. All *Sad* genes except *Sad7* possess a conserved gene structure

Sad7 and its close homologues have not only experienced GC_3 enrichment, but also significant intron losses compared to the tandem paralogous group. The gene structure of the monocot SCPL Clade1A sequences varies tremendously (Figure 5.31), ranging from a 14 exon structure, as for a typical SCP gene, to a single exon, as seen in oat *Sad7* and BRADI3G21550. In contrast, all other *Sad* genes share a conserved gene structure with their homologues.

A complete loss of introns is indicative of a retroposition event. Alternatively, unequal crossing over between paralogues leading to extensive loss of introns may have occurred (Park and Choi, 2010; Zhang et al., 2011). Retroposition events are defined as possessing at least two of three hallmark characteristics: (1) extensive loss of introns compared to the donor sequence, (2) remnant of the poly-(A) tail within 1kb from the 3′ end of the gene, and (3) target site duplication sequences marking the boundaries of retroposition (Zhang et al., 2005). *Sad7* possesses only the first of these characteristics, although the likely age of such an event (before the *Poaceae* radiation 52 mya) means that the two other characteristics would be unlikely to have been retained. However, the tandem arrangement of the *Sad7* homologues lessens the plausability of the retroposition scenario. Furthermore, *Sad7* orthologues were found to possess the three main features of tandem duplicates: simple gene structure, high GC_3 content, and elevated evolutionary rates (*dS*) (Jiang et al., 2013; Wang, 2013)).

Growing evidence also exists of accelarated intron loss and intron divergence in duplicated genes (Zhang et al., 2011). Furthermore, retroposition followed by intron gain is extremely rare, as it involves one retroposition event, one reverse-splicing event, and one recombination (Roy and Irimia, 2009). Therefore, *Sad7* orthologues have most likely experienced extensive intron loss after

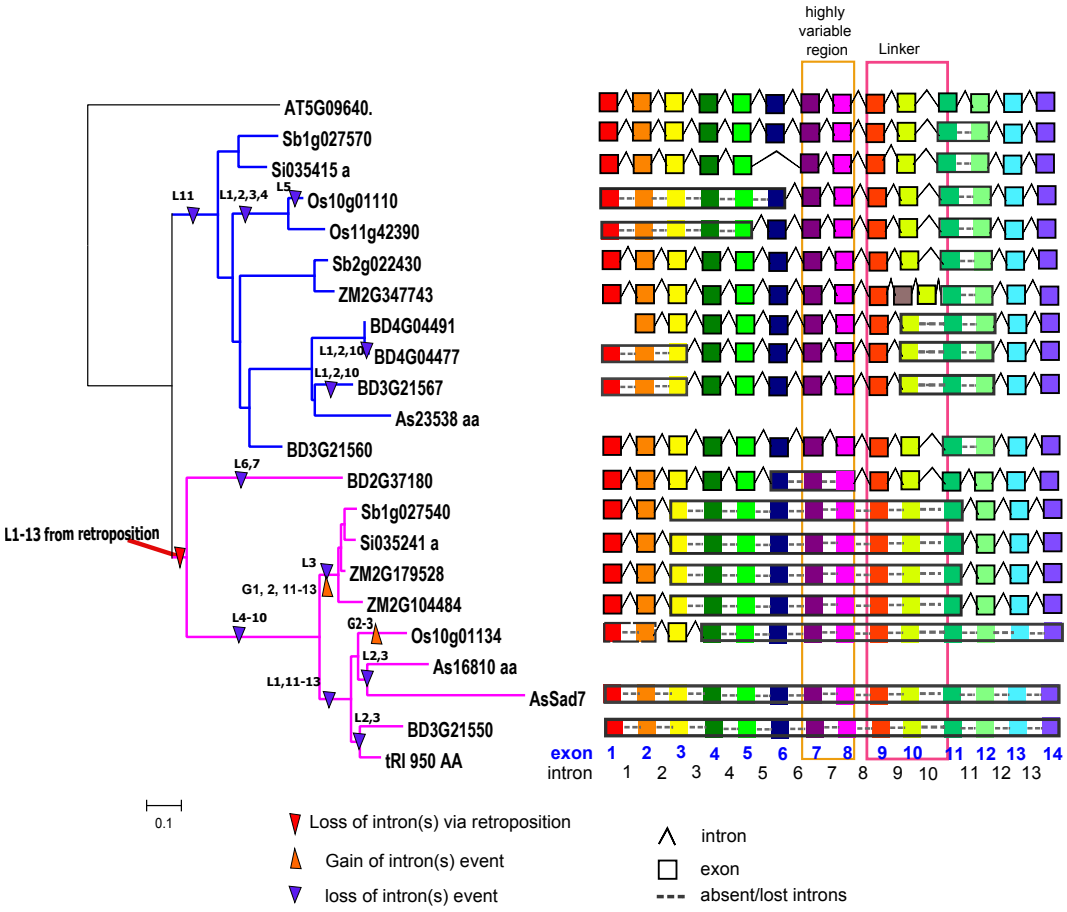


Figure 5.31: The two possible scenarios of intron loss across the *Sad7* phylogeny. 1) Multiple rounds of intron loss events occurred (indicated as purple triangles) with the lost intron numbers adjacent to them (L). 2) Ancient retroposition (indicated by a red triangle) followed by independent gain of intron events (indicated as orange triangles) with the gained intron numbers adjacent to them (G). Exon and intron numbers are defined in accordance with the gene structure of the outgroup sequence *A. thaliana* AT05G09640. Exons 1-14 are colour-coded.

tandem duplication, potentially leading to sequence and functional divergence (Matsuno et al., 2009; Park and Choi, 2010) via changes in expression level.

5.8.7. Limited detection of positive selection in *Sad* gene evolution

Although evidence of positive selection in the evolution of avenacin biosynthesis was expected prior to analysis, relatively few positively selected branches were identified. Most instances involved branches subsequent to ancient duplications that marked the emergence of secondary metabolism (B1 in Figure 5.8, B1 in Figure 5.14, B1 in Figure 5.24 and B1 in Figure 5.29). However, in a small amount of cases positive selection was detected in long branches, either 1) leading to the tips of *Sad* genes, 2) marking functional divergence of *Sad* gene homologues, or 3) immediately following tandem duplications.

Pairwise comparison of loci separated by branches that were not detected to be positively selected showed on several occasions that *dS* values were very high between the gene pairs, resulting in very low *dN/dS* values. These high *dS* values may have been due to saturation in synonymous mutations (Vanneste et al., 2013), affecting internodal ancestral sequence reconstruction, and further reflected as tree topology incongruity between amino acid and codon trees.

Homologues of *Sad2, Sad7, Sad9*, and *Sad10* were found to possess variable GC_3 content, and were likely to have experienced a fast evolutionary rate (*dS*) (Serres-Giardi et al., 2012; Wang, 2013). In at least some of these cases, the variable GC_3 content may have led both to observed phylogenetic inconsistencies between amino acid and codon sequence-derived datasets, and to decreased power to detect positively selected branches. However, further analysis would be required to assess the potential contribution of gBCG and selection for high or low GC/GC_3 content on the elevated dS amongst *Sad* gene homologues before deciding between these competing hypotheses.

Chapter 6 - Investigating the evolution of the avenacin gene cluster

6.1.Introduction

From the phylogenetic studies carried out in Chapter 4 (the analyses of *Sad* gene families amongst oat species) and Chapter 5 (*Sad* gene families amongst monocot species), two sets of sequence data have been generated and analysed, describing the phylogenetic relationships of the cloned *A. strigosa* S75 *Sad* genes to their closely and distantly related homologues. Further analyses of these datasets are reported in this chapter to make inferences about the evolution of the *Sad* genes. In particular, the analyses are focused on 1) the key duplication events marking the emergence of the *Sad* genes and 2) changes in GC content during the evolution of the *Sad* genes.

6.1.1.Timeline of *Sad* gene evolution

With regard to the identification of duplication events, through comparing the five *Sad* gene trees individually with the monocot species tree, the dates and the sequential order of several events key to formation of the avenacin biosynthetic pathway could be inferred. It was previously found that the ρ-whole genome duplication (ρ-WGD) event specific to the monocots was a key event for the emergence of penta-cyclic terpene synthases (Xue et al., 2012). Through queries of the phylogenetic trees of *Sad1, 2, 7, 9,* and *10*, some of the evolutionary events leading to *Sad* gene specialization and avenacin gene cluster formation may be inferred by examining both shared histories of whole genome duplication (WGD) and gene family-specific small scale duplication (SSD).

Calibrating the gene trees of *Sad* genes using WGD and important divergence events

Whole genome duplication events are common evolutionary processes that occur repeatedly in eukaryotic lineages (Murat et al., 2012). Comparative studies of increasing numbers of plant genome sequences have enabled the identification of very ancient WGD events (for example, the ancestral seed plant WGD ζ 349-347 mya, and the ancestral angiosperm WGD ε 234-236 mya) that occurred before the divergence of monocot and dicot plants (130-190 mya), showing that higher land plants are paleopolyploids (Figure 6.1) (Jiao et al., 2011).
Reconstruction of syntenic blocks in *S. bicolor* and *O. sativa* has enabled the identification of an additional ancient duplication event (σ-WGD) in the monocot lineage, dated approximately 130 mya (Figure 6.1) (Tang et al., 2010). Comparative genomics analyses have also determined that the ancestral monocot karyotype (AKG) contained only five chromosomes (Murat et al., 2012). Furthermore, the AKG is thought to have undergone one further WGD (the ρ-WGD $\sim$ 70 mya) event, followed by four chromosomal fissions and two chromosomal fusions, resulting in the intermediate ancestral grass with twelve chromosome pairs. *O. sativa* has retained the chromosome structure of this intermediate ancestral grass whereas multiple genome shuffling events have occurred independently in other lineages, resulting in the different chromosome numbers of modern grass species (Figure 6.2) (Murat et al., 2012; Schnable et al., 2012). It is further believed that nested chromosome fusions led to reductions in chromosome numbers in the *Brachypodium* and *Panicoideae* ancestors and that genomic reshuffling events in general have played a key role in grass speciation (Figure 6.2) (Murat et al., 2012). Furthermore, a recent WGD in *Z. mays* has occurred after its divergence from *S. bicolor* ($\sim$ 13 mya) (Bennetzen et al., 2012).
Characterization of paleo-WGD gene pairs (homeologues) has been based on synteny mapping and age distribution analyses of duplication events (Schnable et al., 2012; Tang et al., 2010). The average rates of synonymous substitutions per synonymous site of syntenic blocks (*Ks*, referred as *dS* from now on) were measured between WGD homeologues with analyses of the subsequent *dS* distributions leading to dates for the paleo-WGD events of 130 mya (σ-WGD) and 70 mya (ρ-WGD) (Figure 6.1) (Jiao et al., 2011; Tang et al., 2010).
Although *dS* between syntenic blocks is frequently used to date evolutionary events through its presumed neutral occurence, it should be used with caution as it can be sensitive to sequence alignment length, saturation effects, selection on transition bias and codon usage (Vanneste et al., 2013). In this chapter, syntenic mapping of the *Sad* gene homologues was carried out across several monocot

species, using common WGD events to calibrate the five individual *Sad* gene family trees. Subsequent inferences on the timeline of avenacin biosynthesis emergence have provided insights into whether stepwise recruitment or modular evolution marked the birth of triterpene metabolism.

6.1.2. The impact on gene specialisation by changes in GC content

The second focus of the analysis presented here, that of GC content alteration, is an emerging theme in plant evolutionary biology. Among plants, it was first described in rice, maize and barley (Carels and Bernardi, 2000) that the GC distribution of genes is bimodal, with both GC-rich and GC-poor genes within each genome. This observation has subsequently been validated as the general trend amongst monocot species (Serres-Giardi et al., 2012; Tatarinova et al., 2010) with GC_3 the most variable component of the GC measure (Figure 6.3). The GC landscape of a gene can influence transcriptional regulation by providing targets for DNA methylation, leading to changes in nucleosome occupancy (Shabalina et al., 2013). Besides contributing to the codon bias of genes, the GC landscape also plays an important role in translational regulation via shaping the mRNA secondary structure, and the interaction of the mRNA with miRNAs, ribosomal proteins and other components of the translational machinery (Shabalina et al., 2013).

Recently, the relationship between GC_3 content, gene expression and DNA methylation levels has been studied in detail in *O. sativa* and *A. thaliana* (Tatarinova et al., 2013). It has been further shown that the intron-exon structure and GC landscape of a gene collectively affect its DNA methylation, expression level and alternative splicing pattern (Amit et al., 2012; Gelfman et al., 2013; Tatarinova et al., 2013). Rice genes do not only display a bimodal GC_3 but also gene-body methylation level distribution (Figure 6.3) (Takuno and Gaut, 2013; Tatarinova et al., 2013). GC_3-rich genes in rice tend to have lesser gene-body methylation levels, wherease GC_3-poor genes have higher gene-body methylation levels (Tatarinova et al., 2013). Of note, GC_3-rich genes in rice have more variable expression levels across tissues compared to the GC_3-poor genes, suggesting that GC_3 may contribute to differential expression patterns of genes in rice (Tatarinova et al., 2013).

In contrast, *A. thaliana* genes exhibit a unimodal GC_3 distribution and a bimodal gene-body methylation distribution (Tatarinova et al., 2013). It was again found that genes with high GC_3 content ($GC_3 = 0.5$) in *A. thaliana* exhibit differential methylation patterns in shoots and roots compared to

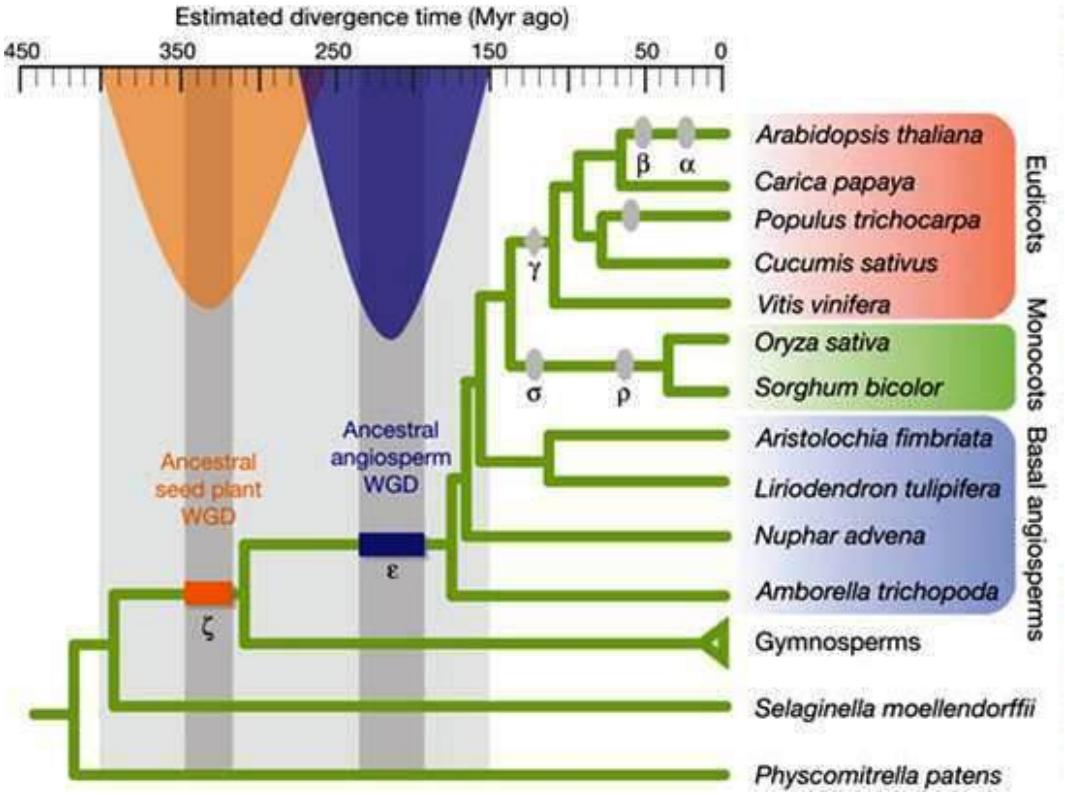


Figure 6.1: The ancient whole genome duplication events of land plants identified by phylogenomic evidence. Reproduced from Jiao et al. (2011).

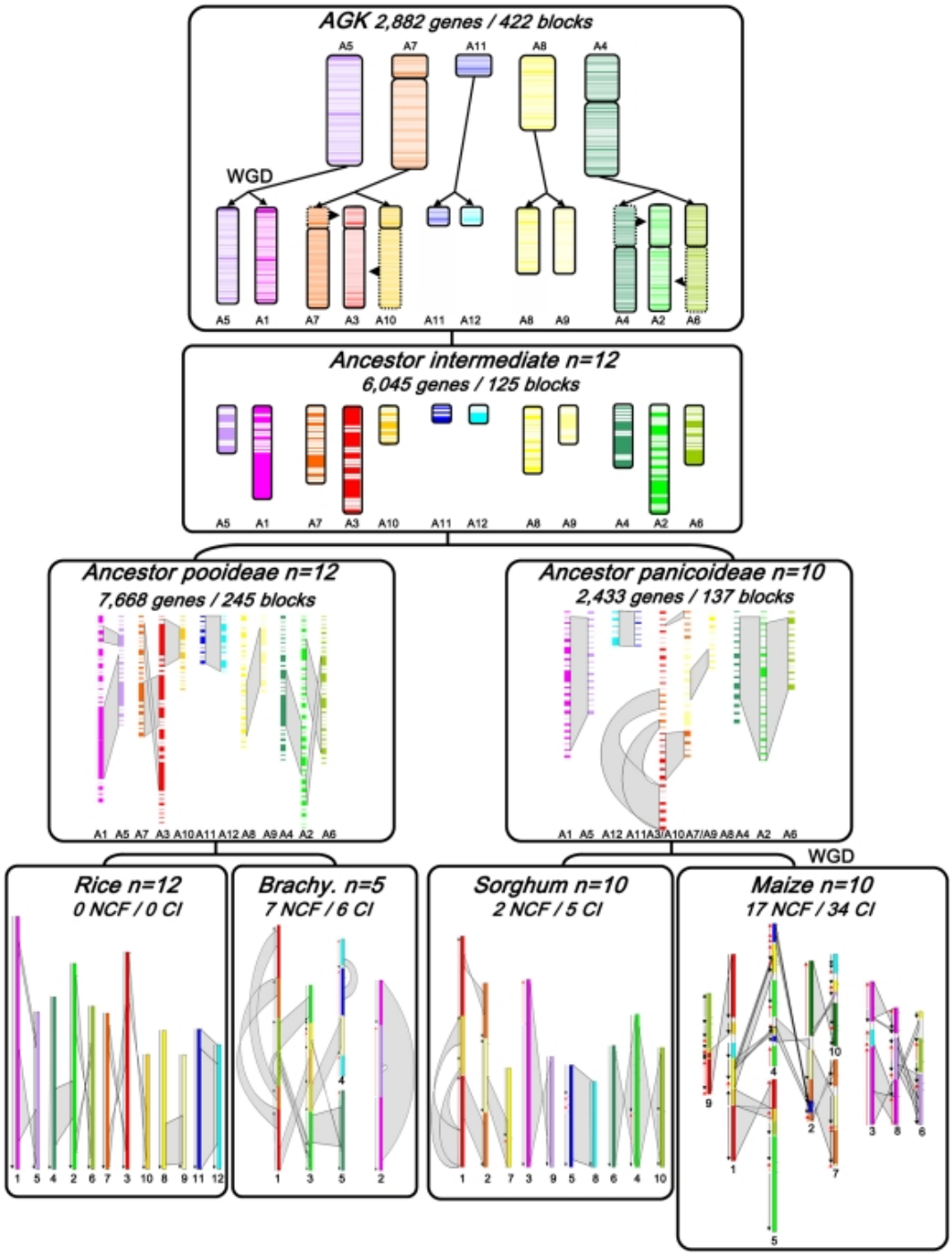


Figure 6.2: The ancestral grass karyotype (AKG) and the genome rearrangement events leading to the modern chromosomal structure of rice, *Brachypodium, Sorghum* and maize. The chromosomes (A5, A7, A11, A8 and A4) of the AKG and their syntenic blocks are color-coded. WGD - whole genome duplication or polyploidisation events. n - number of chromosomes. The number of nested chromosome fusions (NCFs) and chromosome inversions (CIs) are indicated. Reproduced from Murat et al. (2010).

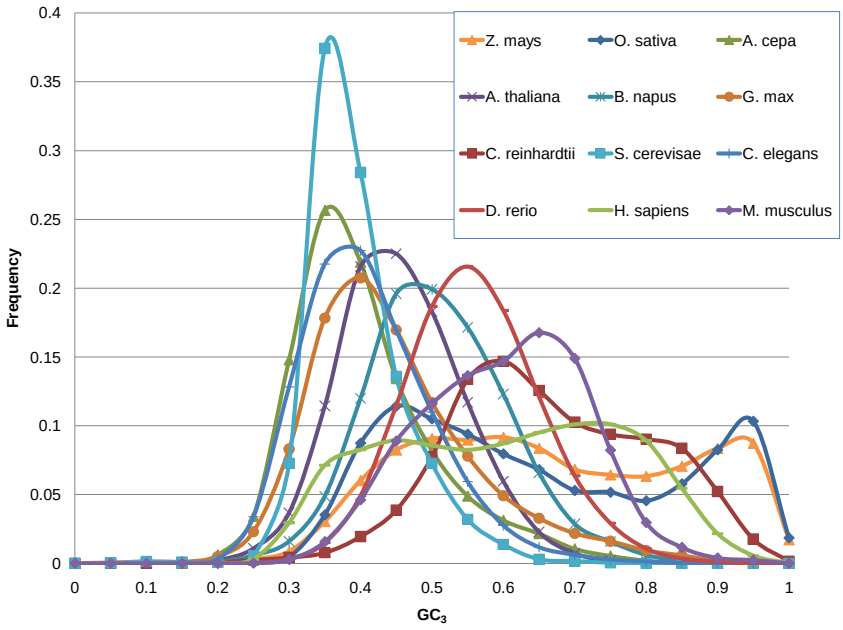


Figure 6.3: Genomic GC_3 distributions. Reproduced from Tatarinova et al. (2010)

GC-poor genes ($GC_3 < 0.4$), highlighting the contribution of methylation to tissue-specific expression (Tatarinova et al., 2013).

6.1.3.Selection on synonymous sites for GC_3 content

In Chapter 5, large variations in GC content within the phylogenies of *Sad1, 2* and *7* were uncovered. These GC changes were strongest at the third codon positions, which are 'silent' in terms of the amino acid sequences but are key players in the regulation of gene expression (Amit et al., 2012; Gelfman et al., 2013; Tatarinova et al., 2013). It is possible that selection on GC_3 landscapes of the *Sad* gene homologues has led to divergence of tissue-specific expression from their paralogues. Another, not mutually exclusive, explanation is that such changes in GC landscapes are contributed to by GC-biased gene conversion, a selectively neutral process. GC-biased gene conversion (gBCG) is a DNA mismatch repair mechanism that occurs frequently during recombination in land plants and that is believed to be the main driver behind the bimodal gene GC content observed in monocots (Serres-Giardi et al., 2012; Tatarinova et al., 2010). Of note, gBCG is a feed-forward mechanism that occurs more frequently in GC-rich genomic regions, further increasing the GC content of these regions (Tatarinova et al., 2010). In plants, a clear positive correlation between GC_3 content and recombination freqency at megabase scales in the genome has been observed (Serres-Giardi et al., 2012), potentially leading to high GC genomic regions enriched with gBCG. Therefore, the GC changes observed amongst *Sad* gene homologues may be due to their movement into or out of gBCG-enriched regions via gene duplication or genome reorganisation events.

As discussed in Chapter 5, high levels of non-synonymous nucleotide changes observed in the evolution of the *Sad* genes appear to be associated with changes in GC content, potentially affecting the detection of signals of positive selection by PAML branch-site tests. Branch-site tests are ultimately based on the *dN/dS* ratio, and assume that synonymous changes are selectively neutral, driven by the natural rate of mutation under the molecular clock (Yang, 2006). If the elevated synonymous changes that have been observed in the *Sad* gene phylogenetic studies are attributable to selection for changes in genic GC contents for differential gene expression (i.e *dS* levels are inflated), PAML selection tests may be unable to detect signals of positive selection. Unfortunately, systematic approaches to detect signals of selection for non-synonymous substitutions in the presence of selection for synonymous substitutions have not yet been developed.

In this chapter, *Sad2* homologues are used as a case study to investigate whether

the synonymous changes that have occurred in these genes are likely to be due to selection for gene GC landscape changes or simply consequences of the genes relocating into high or low GC isochores. The analysis was then extended to *Sad1* homologues to determine whether they also experienced a similar mode of GC content evolution.

6.2.Materials and methods

6.2.1.Syntenic mapping of paleo-WGD gene pairs

The phylogenetic trees in Chapter 5 for the *OSCs* (Figure 5.26), *CYP51s* (Figure 5.20), *SCPLs Clade1A* (Figure 5.5a), *Class I OMTs* (Figure 5.10), and *Group L UGTs* (Figure 5.15) using RAxML 7.0.4 (Stamatakis, 2006) were used as the framework for the analysis. Potential paralogous gene pairs of rice, *Sorghum* and *Brachypodium* that originated from the ρ-WGD event were first identified by syntentic mapping using the SyMap webserver (Soderlund et al., 2011) (http://www.symapdb.org/). Syntenic ρ-WGD paralogous pairs were further validated by estimating pairwise dS values using the PAML 4.5 CODEML (runmode -2) (Yang, 2007) software. The estimated dS values were then compared to the reference dS values for the ρ-WGD events, as reported by Tang et al. (2010) and Schnable et al. (2012).

6.2.2.Surveying GC landscapes of *Sad* genes

For each *Sad* gene, the overall GC, GC_1, GC_2, and GC_3 values of exons and the total GC content of introns were measured as described in Chapter 5 using the Seqinr package in R (Charif and Lobry, 2007; R Core Team, 2012).
GC contents were measured in a sliding window of 100 bp, with steps of 1bp along the gene. To determine how genic GC content varies with the surrounding environment, the GC content of the genomic region 5 kb upstream downstream of the gene of interest was also measured. The genomic region of the entire gene cluster was also analysed using the GCprofile webserver for isochore(s) identification (Gao and Zhang, 2006). The GC content along the *A. strigosa* S75 avenacin gene cluster was measured in a non-overlapping sliding window of 300 bp to assess the effects of GC content surrounding the genic regions of *Sad1, 2, 7, 9* and *10* on their genic GC content.

6.2.3. Evaluation of four-fold degenerate sites in GC content changes

Conserved four-fold degenerate (4D) sites were identified in the *Sad2* homologues using the pairwise amino acid alignment of the rice *CYP51G1* (under purifying selection) and the *CYP51H* genes. The GC_3 values of the pairwise conserved 4D sites (defined as GC_4) of the *CYP51G1* and the corresponding *CYP51H* sequences were measured. The association of the pairwise *dS* values (obtained from PAML 4.5 (Yang, 2007)) and the difference in GC_4 values between the *CYP51G1* and the corresponding *CYP51H* genes were evaluated.

6.3. Results

6.3.1. Identification of ρ-WGD paralogous gene pairs

Studying the genomic context of these two *OSC* ρ-WGD pairs, the orthologous tandem *OSC* array present on rice chromosome 2 and *S. bicolor* chromosome 4 and their ρ-WGD duplicates were again identified (Xue et al., 2012). Consistent with previous findings (Xue et al., 2012), the ρ-WGD homeologues in rice have been lost (Figure 6.4b). The *OSC* phylogeny indicates that an ancient tandem duplication that occurred before the ρ-WGD event has given rise to triterpene biosynthestic *OSC* genes that functionally diverged from *CAS1* (Figure 6.4a). A subsequent tandem duplication after the ρ-WGD event of the triterpene biosynthetic *OSC* genes gave rise to LOC_Os02g04730 and LOC_Os02g04750. Similarly in *S. bicolor*, Sb10g029150/Sb10g029175 and Sb04g03210/Sb04g03300 have arisen from independent tandem duplications following the ρ-WGD event that gave rise to their ancestors.

It may be speculated that homeologous *CAS* duplicates generated by the ρ-WGD event have been lost in both rice and sorghum due to selective disadvantage from dosage effects. It could also be speculated that the tandemly duplicated *OSC* genes were not required for primary sterol biosynthesis and thus were retained by selection for neo-functionalization or genetic drift (in physical proximity to cycloartenol synthase). In the *OSC* tree, the *Sad1* clade does not follow the species phylogeny (Figure 5.26, 6.4a), suggesting that specialization of *Sad1* is likely to have occurred very recently and may be repeated in *S. bicolor, S. italica* and maize.

The ρ-WGD duplicate pair within the *CYP51* tree was found to be implicated in the divergence of the *CYP51* genes from the primary metabolic *CYP51G1* genes (Table 6.1, Figure 6.5). Syntenic mapping and phylogenetic analysis showed that

Gene family	WGD homeologues	ρ block	Pairwise dS (gene pair)	Median dS (ρ block)
OSC	Sb04g03210-Sb10g029150	ρ3	0.99	0.98
OSC	Sb04g03300-Sb10g029175	ρ3	0.91	0.98
CYP51	Sb05g019590-Sb08g002250	ρ8	0.84	0.51
OMT	Sb04g036890-Sb10g027360	ρ3	1.0	0.98
OMT	Sb05g019040-Sb08g005125	ρ8	10.80	0.51
UGT75	LOC_Os02g10880-LOC_Os06g39330	ρ3	1.69	0.98

Table 6.1: Summary of ρ-WGD duplicates. Summary of the gene pairs in the phylogenetic trees of *OSCs* (Figure 5.26), *CYP51s* (Figure 5.20), *SCPLs Clade1A* (Figure 5.5a), Class I *OMTs* (Figure 5.10) and Group L *UGTs* (Figure 5.15) that are likely to originate from the ρ-WGD event. The median dS (Tang et al., 2010) was used as a reference with which to compare to the dS values estimated for *Sad* gene pairs.

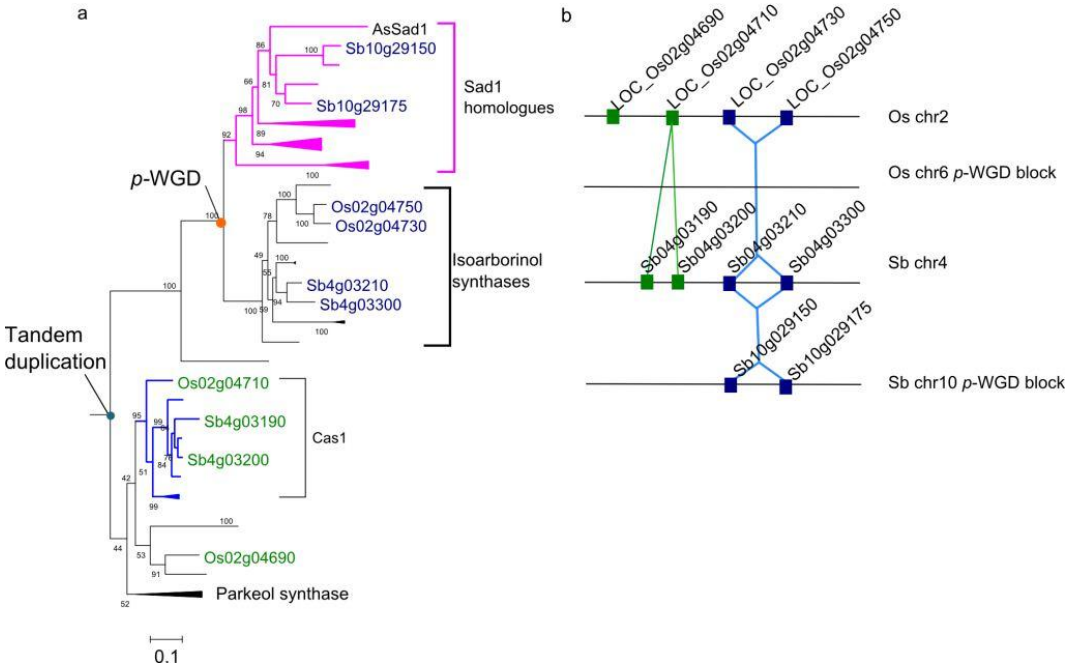


Figure 6.4: The paralogous *OSC* gene pairs originate from the ρ-WGD event. a) A simplified RAxML *OSC* amino acid tree (Figure 5.26a) with the key tandem duplication event and the ρ-WGD event indicated. *OSC* clades are labelled according to published annotations (Xue et al., 2012). b) Collinearity of ρ-WGD blocks of *OSC* in rice and *S. bicolor*. The orthologues, ρ-WGD duplicated genes and genes sharing a recent common ancestry are connected.

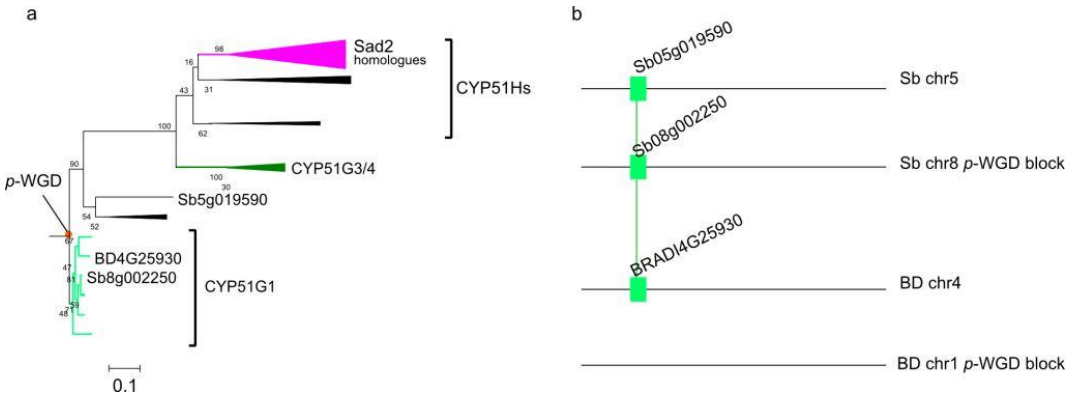


Figure 6.5: The paralogous *CYP51* gene pair originated from the ρ-WGD event. a) A simplified RAxML *CYP51* amino acid tree (Figure 5.20a) with the ρ-WGD event indicated. b) Collinearity of ρ-WGD blocks of *CYP51* in *S. bicolor* and *B. distachyon*. The orthologues and ρ-WGD duplicated genes are connected.

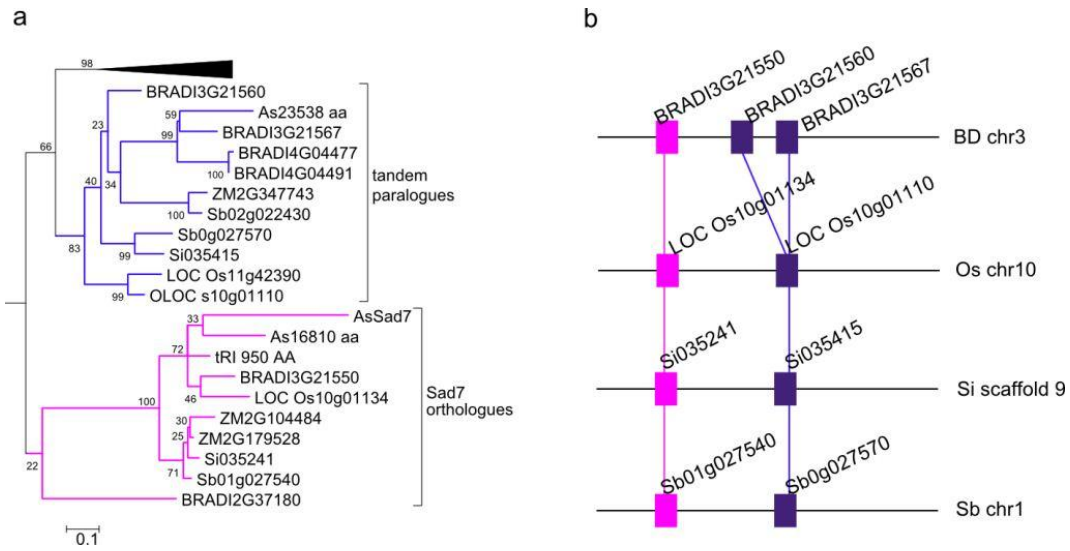


Figure 6.6: The paralogous *SCPL* gene pair originated before the radiation of the *Poaceae*. a) A simplified RAxML *SCPL1* amino acid tree (Figure 5.5) showing the *Sad7* orthologues and their tandem gene duplicates. b) Collinearity of *SCPL1* orthologues in *B. distachyon, O. sativa, S. italica* and *S. bicolor*. The orthologues are connected.

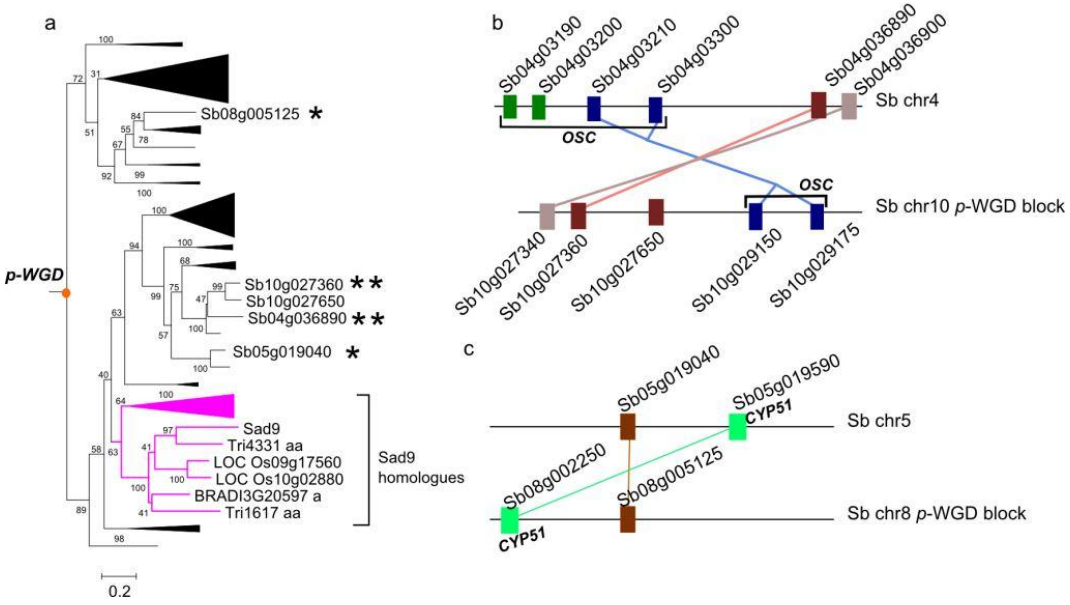


Figure 6.7: The two paralogous Class I *OMT* gene pairs likely to have originated from the ρ-WGD event. a) A simplified RAxML Class I *OMT* amino acid tree (Figure 5.10a) showing the phylogeny of the ρ-WGD paralogous gene pairs, Sb04g036890/Sb10g27360 (* *) and Sb05g019040/Sb08g005125 (*), which are highlighted with asterisks. The likely ρ-WGD event is indicated. b and c) Collinearity of ρ-WGD blocks of Class I *OMT* genes in *S. bicolor*. The potential ρ-WGD duplicated genes are connected. The *Sad1* and *Sad2* ρ-WGD gene pairs sharing the same syntenic blocks are shown.

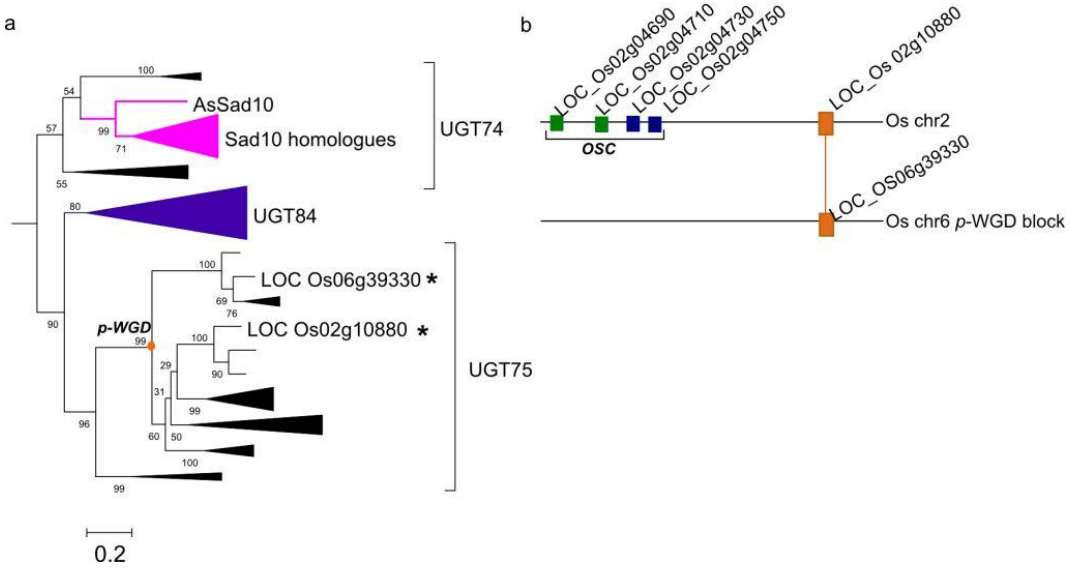


Figure 6.8: The paralogous Group L *UGT* gene pair likely to have originated from the ρ-WGD event. a) A simplified RAxML Group L *UGT* amino acid tree (Figure 5.15a) showing the phylogeny of the UGT75 ρ-WGD paralogous gene pair with *UGT84s*, *UGT74s*, *Sad10* and its closely related homologues. The ρ-WGD homeologues within the UGT75 clade are marked with an asterisk. The node denoting the ρ-WGD event is indicated b) Collinearity of ρ-WGD blocks of the UGT75 gene pair in rice. The rice *OSC* tandem array sharing the same ρ-WGD block with LOC_Os02g10880 is also indicated. The ρ-WGD duplicated genes are connected.

Sb05g19590 is derived from Sb08g002250 via the ρ-WGD event (Figure 6.5a and b). Orthologous to Sb08g002250, BRADI4G25930 on *B. distachyon* chromosome 4, however, appears to have lost its ρ-WGD homeologues in the syntenic region of chromosome 1 (Figure 6.5b).

The RAxML SCPL Clade1A amino acid tree (Figure 5.5a), showed that the *Sad7* orthologue in each of the surveyed monocot genomes has a tandem paralogue (Figure 6.6). This tandem gene pair was likely to have arisen before the divergence of the *Poaceae* because of its presence in all analysed monocot species with the exception of maize. However, no ρ-WGD duplicates could be identified in the SCPL Clade1A tree. Therefore, it was likely that such a tandem array formed after the ρ-WGD event. Interestingly, the pairwise *dS* values between these tandem duplicates are greater than 80, which may be attributable to the *Sad7* orthologues experiencing increased GC content evolution and radical changes in gene structure (discussed in Chapter 5).

Two potential ρ-WGD gene pairs were identified within the Class I OMT tree (Table 6.1, Figure 6.7a and b). The Sb04g036890/Sb10g027360 gene pair was found to share the same syntenic block with the *Sad1* ρ-WGD duplicates (Figure 6.7b). Their tandem duplicates, Sb04g036900 and Sb10g027340, although not included in the phylogeny due to their short lengths, also share a high level of sequence similarity. Sb10g027650 is likely to have derived from a recent tandem duplication from the ancestor of Sb10g027360. However, Sb04g036890 and Sb10g027360 cluster within the *Panicoideae*-specific clade in the Class I OMT tree (clade denoted 'P' in Figure 5.10a), suggesting that these two genes were unlikely to have arisen via the ρ-WGD event despite a consistent *dS* value within the relevant syntenic block unless both genes have undergone gene conversions with a recently emerged *OMT* paralogue (i.e. Sb10g027650), leading to their appearance to be closely related in the phylogenetic tree. Instead, a recent segmental duplication within the *Panicoideae* has taken place and gave rise to these genes.

The other potential ρ-WGD gene pair Sb08g005125/Sb05g019040 shares the same syntenic block with the *CYP51* ρ-WGD pair (Table 6.1, Figure 6.7a and c). The high pairwise *dS* value of this gene pair may be due to extensive sequence diversification. If Sb08g005125/Sb05g019040 is indeed a ρ-WGD gene pair, the ρ-WGD event would be implicated in the early divergence events of Class I monocot OMTs. Most importantly, there were no ρ-WGD pairs in the OMT homologue clade, consisting of the most closely related genes to *Sad9*. In addition to the fact that closely related *Sad9* homologues were only present in the species belonging to the *Bambusoideae, Ehrhartoideae*, and *Poodieae* (BEP) subfamilies (Figure 6.7a), this finding further supports the hypothesis that OMT

sequences involved in triterpene biosynthesis were likely to have emerged after *Sad1*- and *Sad2*- like sequences have arisen.

AsSad10 was shown in the Group L UGT tree to be distantly related to all of its closest homologues, located at the base of the clade (Figure 6.8a). The paucity of *Sad10*-like sequences amongst the identified monocot UGT74s in the analysis suggests that *Sad10*-like genes may be *Avena*-specific. The two clades encompassing the *Sad10* clade do not appear to possess any ρ-WGD related genes. However, a potential ρ-WGD pair of rice genes in the UGT75 clade, LOC_Os02g10880 and LOC_Os06g39330, was identified. LOC_Os02g10880 is approximately 3 MB from the OSC tandem array LOC_Os02g04690-730, located on the same ρ-WGD block (Figure 6.8b).

6.3.2.Sequence of *Sad* gene emergence

The timings of the various WGD and tandem duplications in the five *Sad* gene trees suggest that avenacin biosynthesis was likely to have evolved via duplication of the core *Sad1-Sad2* module and extended through stepwise recruitment of *Sad7* and *Sad9* (Figure 6.9). In the broad evolution of avenacin biosynthesis, the emergence of '*Sad1*'- and '*Sad2*'-like genes from primary sterol biosynthesis is likely to have occurred through the ρ-WGD event ($\sim$ 70 mya). The tandem duplication event that gave rise to the *Sad7*-like sequence was likely to have occurred before the divergence of the *Poaceae* but after the ρ-WGD at approximately 50 to 40 mya. *Sad9*-like genes were likely to have emerged before the BEP clade divergence ($\sim$ 50-35 mya). By that time, the basic gene set for triterpene synthesis could be found in the ancestral BEP monocot genome. The emergence of *UGT74s* was likely to have occurred before the species radiation of *Poaceae* but the ancestral gene copy of *Sad10* appears to be oat-specific. Furthermore, genuinely closely related *Sad* gene orthologues cannot be found beyond *Avena* species, which diverged from other grasses less than 20 million years ago (Wu and Ge, 2012). The phylogenetic studies of *Sad* genes in Chapter 5 also show that gene clustering of *Sad* gene homologues has not been found to occur in any cereals other than oats. Therefore, it can be concluded that *Sad* gene specialisation for avenacin biosynthesis and the formation of the gene cluster are recent events.

6.3.3.GC landscape of the *Sad* genes

The GC and GC_3 contents of the oat *Sad1*s are on average 0.44 and 0.41 (Table 6.2), which fall within the GC-poor class of monocot genes (Carels and Bernardi,

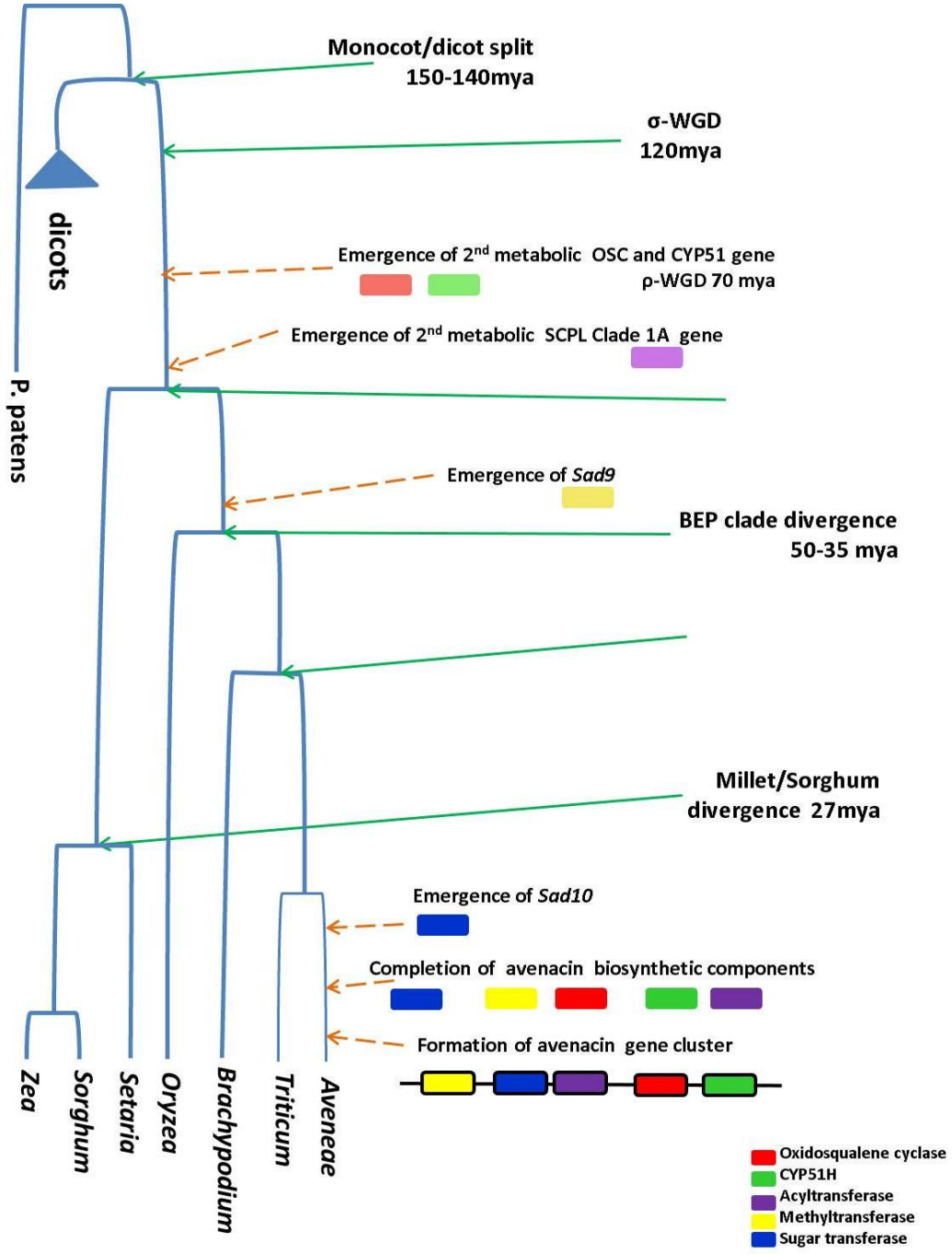


Figure 6.9: The proposed timeline of *Sad* gene emergence. The key speciation and WGD events of monocots (Bennetzen et al., 2012; Christin et al., 2008; DHont et al., 2012; Jiao et al., 2011; Tang et al., 2010; Wu and Ge, 2012) were aligned with the key duplication events giving rise to *Sad1, 2, 7, 9* and *10*. The timing of the formation of the *Sad* genes was likely to be after the divergence of *Aveneae* from other grasses (~ 18-13 mya).

2000). A GC content plot, using a sliding window of 100 bp along the genomic sequences of oat *Sad1* homologues (Appendix Figure 6.1), shows that the homologous oat *Sad1* genes contain a differential gene GC structure (Amit et al., 2012) (low GC within exons and even lower GC within introns). GC analysis further shows that exon 1 of the oat *Sad1* genes contains the highest GC content (GC/GC_3 content is 0.60/0.75) amongst the 18 exons, two fold higher than all other exons (Appendix Table 6.1). A GC plot of the *Sad1* region shows that GC content peaks at the translation start site (Appendix Figure 6.1). Exon 14 of the *Sad1* homologues is shown to contain the lowest GC content amongst all exons (Appendix Table 6.1). Most exons possess a higher GC gradient at the 5′ end and dropping gradually to the 3′ end, suggesting that the *Sad1* GC landscape plays a role in defining splice sites.

Oat *Sad2* homologues also possess a low GC content (average GC/GC_3 = 0.44/0.47) and display a differential gene GC structure (Table 6.3). Unlike the *Sad1* homologues, the GC contents of both exons of *Sad2*s are highly similar (Table 6.2 in the Appendix). The GC plot along the *Sad2* homologues reveals the GC content peaks to be located at the 5′ end of exon1 and the middle of exon2 (Appendix Figure 6.2). Similar to *Sad1*, the GC content of *Sad2* exons is higher at the 5′ end and drops rapidly at the 3′ end, potentially defining the exon junctions (Appendix Figure 6.2).

To determine whether both *Sad1* and *Sad2* are located in low GC genomic regions, the GC content along the avenacin gene cluster of *A. strigosa* was measured with a non-overlapping sliding window of 300 bp. The genic regions of *Sad10* and *Sad7* appear to be located at GC peaks bounded by lower GC regions. *Sad1* appears to be within a GC valley bounded by higher GC regions (Figure 6.10). *Sad9* and *Sad2* did not appear to contain GC contents different from their surrounding regions. A GCprofile analysis (Gao and Zhang, 2006) revealed that the avenacin gene cluster consists of two isochores. *Sad9* and *Sad10* are located within the first isochore (1-85554 bp), possessing a higher GC content (approximately 0.457) (Figure 6.11). In contrast, *Sad7, 1*, and *2* are located within a second isochore (85554-243247 bp) with lower GC content (approximately 0.409). This isochore separation of the avenacin biosynthetic genes may suggest the presence of different modes of regulation of expression of the *Sad* genes. Alternatively, the GC content surrounding *Sad7, 1*, and *2* may have been selected to be low to supress gBCG that may lead to the loss of a favourable allelic combination (Tatarinova et al., 2010). An alternative explanation would be that the avenacin gene cluster has been formed by a segmental rearrangement events that has merged the two isochores, one of them containing *Sad1, 2* and *7* with the other containing *Sad9* and *10*.

Species	Genome	Tissue expression	Sequence length	GC content	GC_1	GC_2	GC_3
A. strigosa S75	As	root	2283	0.43	0.49	0.41	0.40
A. prostrata	Ap	root	2181	0.43	0.48	0.41	0.41
A. damascena	Ad	root	2181	0.43	0.48	0.41	0.40
A. canariensis	Ac	root	2172	0.43	0.48	0.40	0.42
A. longiglums	Al	root	2166	0.43	0.48	0.40	0.41
A. pilosa	Cp	leaf	2181	0.42	0.48	0.41	0.39
A. pilosa	Cp	root	2286	0.43	0.49	0.41	0.39
A. clauda	Cp	leaf	2235	0.42	0.46	0.42	0.39
A. clauda	Cp	root	2280	0.43	0.49	0.41	0.39
A. ventricosa	Cv	leaf	2223	0.42	0.47	0.41	0.38
A. ventricosa	Cv	root	2283	0.43	0.49	0.41	0.39
A. sterilis	ACD	root	2259	0.43	0.48	0.41	0.40
A. fatua	ACD	root	2223	0.43	0.48	0.40	0.40

Table 6.2: GC content of the transcripts of oat *Sad1* homologues retrieved from the RT-PCR analysis in Chapter 3. The GC content of exons 1-18, introns 1, 3, 5-17 of the genomic sequences of oat *Sad1* genes are listed in Appendix Table 6.1.

Species	Genome	Tissue expression	Sequence length	GC content	GC_1	GC_2	GC_3
A. strigosa S75	As	root	1383	0.45	0.51	0.38	0.46
A. prostrata	Ap	root	1341	0.45	0.50	0.39	0.46
A. damascena	Ad	root	1299	0.45	0.51	0.39	0.46
A. canariensis	Ac	root	1272	0.45	0.50	0.38	0.47
A. longiglums	Al	root	1302	0.45	0.50	0.38	0.47
A. pilosa	Cp	leaf	1422	0.44	0.49	0.37	0.46
A. pilosa	Cp	root	1341	0.43	0.50	0.36	0.45
A. clauda	Cp	leaf	1428	0.44	0.49	0.37	0.45
A. clauda 1	Cp	root	1323	0.45	0.50	0.38	0.46
A. clauda 2	Cp	root	1374	0.44	0.50	0.37	0.45
A. ventricosa	Cv	leaf	1422	0.44	0.49	0.37	0.46
A. ventricosa	Cv	root	1221	0.44	0.49	0.35	0.48
A. sterilis	ACD	root	1320	0.45	0.50	0.39	0.46
A. fatua	ACD	root	1386	0.45	0.50	0.39	0.46

Table 6.3: GC content of the transcripts of oat *Sad2* homologues retrieved from the RT-PCR analysis in Chapter 3. The GC content of exons 1-2 and intron 1 of the genomic sequences of oat *Sad2* genes are listed in Appendix Table 6.2.

6.3.4.Correlation of GC changes to *dS* value

The rice *CYP51* genes were analysed further to investigate the effect of GC content alteration on the elevated *dS* values previously observed. The *CYP51G1* gene remained under purifying selection for primary sterol metabolism while the other *CYP51* genes were free to diversify to a new function. Therefore, the sequences of rice *CYP51G1* was compared to other rice *CYP51H* genes to investigate how changes in GC content affect pairwise *dS* values in this family (Table 6.4).

Most rice *CYP51* genes are located in a genomic environment with a GC content of approximately 0.42. Of note, the GC_3 contents of these genes are higher than the genomic GC contents (Table 6.4). The rice *CYP51* genes all possess a differential GC structure. The intronic GC content of the rice *CYP51* genes are lower than 0.40, with the exception of LOC_Os07g37980 (intron GC = 0.42). LOC_Os11g32240, LOC_Os07g28110 and LOC_Os07g28160 display a gene GC landscape in which the GC_3 of exon 1 is higher than that of exon 2, with the most drastic difference being observed in LOC_Os11g32240 (25%). In contrast, the GC_3 of both exons of LOC_Os05g34380 and LOC_Os07g37970 are highly similar, while the rest of the *CYP51* genes display a gene GC landscape in which the GC_3 of exon 1 is lower than that of exon 2, with the most drastic difference being observed in LOC_Os05g12040 (40%).

The three genes with the highest deviation from the average GC content of the rice *CYP51*s also have the highest pairwise *dS* values with LOC_Os11g32240 (Table 6.4). To further evaluate whether the rice *CYP51*s are under selective constraints, the pairwise GC_4 (third codon position of the conserved amino acids proline, alanine, threonine, glycine and valine that can be either A, T, G or C) between the *CYP51H* and *CYP51G1* genes were measured (Table 6.4) (Lawrie et al., 2013).

When a gene is under completely 'neutral' selection for its GC content, GC_4 is expected to be 0.5 (Lawrie et al., 2013). Otherwise, the GC_4 value would be expected to be similar to the GC content of the surrounding genomic environment, which is also likely to be evolving under neutral selection. In the rice *CYP51* gene dataset, none of these genes possess a GC_4 of 0.5, nor follow the environmental genomic GC content (with the exception of LOC_Os02g02230 and LOC_Os05g34330), suggesting that the GC landscapes of these genes are under selective contraints (Table 6.3 in the Appendix).

When comparing the pairwise *dS* values to the pairwise GC_4 differences with LOC_Os11g32240 for these rice *CYP51* genes, an association between the two values can be observed (Figure 6.12). In the future, this complex association pattern between GC_4 differences and *dS* values of the rice *CYP51* genes will be

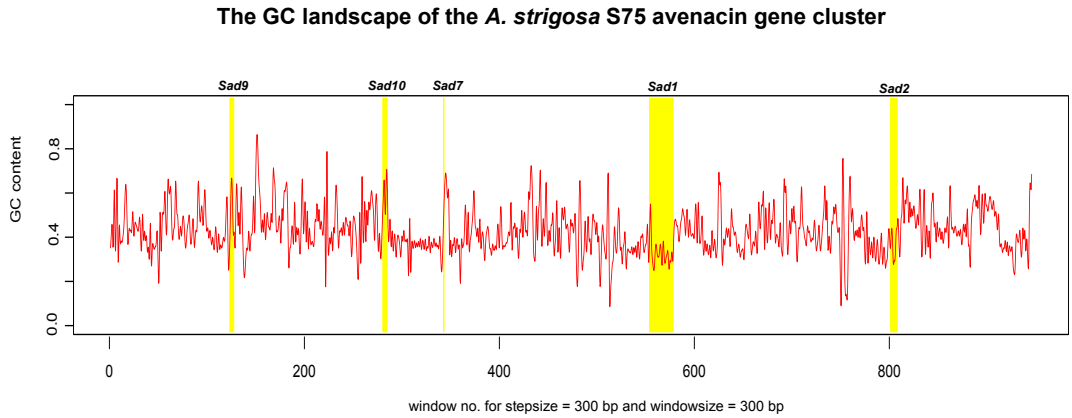


Figure 6.10: GC landscape of the avenacin gene cluster. The GC content of the avenacin gene cluster in *A. strigosa* S75 (280 kb) is meaured using a non-overlapping sliding window of 300 bp along the gene (minimal five sampling points for each *Sad* gene (highlighted in yellow)).

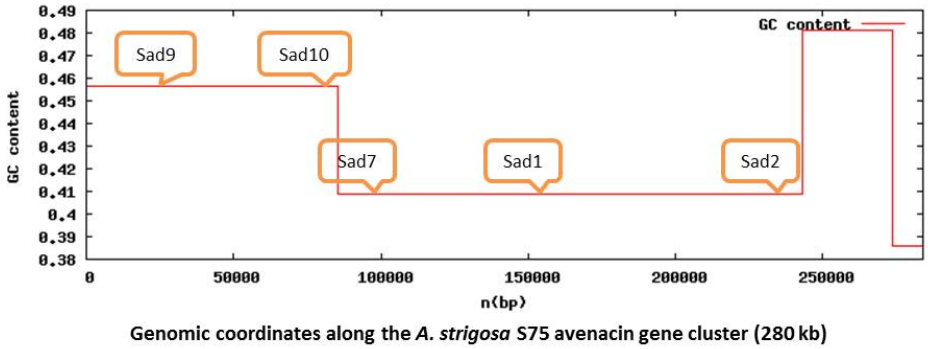


Figure 6.11: Isochore structure of the *A. strigosa* avenacin gene cluster. The two isochores identified by GC-Profile (Gao and Zhang, 2006). The approximate locations of the five *Sad* genes are indicated.

Gene	Genomic GC content	Coding sequence GC	Coding sequence GC_3	Intronic GC	Pairwise GC_4 difference with LOC_Os11g32240	Pairwise dS value with LOC_Os11g32240
LOC_Os02g02230	0.43	0.50	0.59	0.32	0.25	73.6
LOC_Os02g21810	0.44	0.61	0.78	0.32	0.05	31.0
LOC_Os05g12040	0.43	0.60	0.72	0.37	0.07	40.0
LOC_Os05g34325	0.40	0.56	0.68	0.37	0.12	21.6
LOC_Os05g34330	0.43	0.53	0.61	0.30	0.25	11.5
LOC_Os05g34380	0.40	0.45	0.50	0.30	0.27	65.9
LOC_Os07g28110	0.44	0.59	0.74	0.32	0.06	13.7
LOC_Os07g28140	0.36	0.66	0.84	na	0.20	na
LOC_Os07g28160	0.41	0.62	0.79	0.30	0.04	12.0
LOC_Os07g37970	0.45	0.66	0.89	0.31	0.19	97.9
LOC_Os07g37980	0.43	0.60	0.75	0.42	0.08	12.5
LOC_Os11g32240	0.40	0.61	0.84	0.35	na	na

Table 6.4: Summary of GC content analysis of the rice *CYP51* genes. Complete table is in the Appendix (Table 6.3). LOC_Os07g28140 was not included in the PAML analysis because it does not possess exon 2.

fully dissected by a more detailed analysis.

6.4.Discussion

6.4.1.Ancient duplication events leading to triterpene biosynthesis

The phylogenetic studies carried out in this chapter have revealed that the ρ-WGD event (70 mya) was a key duplication event within the *OSC* and *CYP* gene families. It led to new gene clades that were ultimately involved in specialised metabolism of triterpenes in the *Poaceae*. However, the evolution of specialised metabolism for avenacin biosynthesis appears to be a recent event (<20 mya). Surveying the genomes and triterpene biosynthesis profiles of species within the *Pooidiae* will shed further light on how avenacin biosynthesis has emerged from the evolution of the individual triterpene biosynthetic genes.

6.4.2.The effect of the GC landscape changes on gene evolution

The GC analyses carried out here revealed that the changes in GC content during the evolution of *Sad1* and *2* are unlikely to be consequences of gBCG or another mechanism that affects the GC contents uniformly in the corresponding genomic regions. This is because these genes are located in low GC regions and the changes in GC content have mainly taken place in the coding sequences rather than being uniformly distributed across both exons and introns.

Investigation of the 4D-sites indicated a preference for low GC_3 content (Tatarinova et al., 2013) in rice *CYP51H* subfamily genes, the exceptions being LOC_Os07g28140 and LOC_Os07g37970. The marked differences of GC_4 between the rice *CYP51G1* (LOC_Os11g32240) and the *CYP51* genes, LOC_Os02g02230, LOC_Os05g34330, and LOC_Os05g34380, suggest that these genes are likely to possess different methylation patterns (Takuno and Gaut, 2013), which are responsible for tissue-specific expression. Therefore, changing the expression profile of the rice *CYP51* genes may be achieved by altering their GC landscapes. The rice *CYP51* genes with their new expression patterns could then be able to explore different metabolic networks for making novel compounds with new combinations of co-expressing functional partners. Alternatively, the low GC_4 content of these genes may be consequences of selection against gBCG (Tatarinova et al., 2010), preventing these genes from undergoing recombination with other *CYP51* paralogues. To fully address the

selective pressures behind GC changes in synonymous sites of the rice *CYP51* genes, ancestral reconstruction-based analysis would need to be carried out to examine the changes in GC landscape of each of the ancestral intermediates in the evolution of the rice *CYP51* genes.

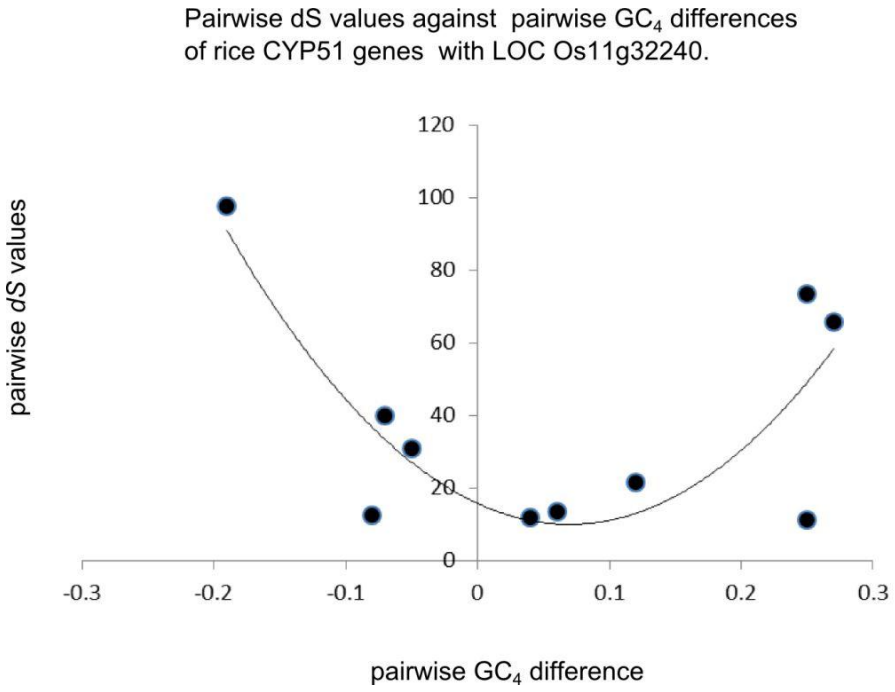


Figure 6.12: Scatter plot of the pairwise *dS* values against GC_4 differences of rice *CYP51* genes with LOC_Os11g32240. The association between *dS* values and GC_4 differences is illustrated by the trend line. The raw data is enclosed in Appendix Table 6.3.

Chapter 7 - General discussion

7.1.Summary of results

Avenacins are antimicrobial triterpene glycosides with important functions in plant defence. The five characterised avenacin biosynthetic genes form part of an operon-like gene cluster in the genome of the diploid oat *A. strigosa* accession S75. This thesis describes the investigation of the evolution of avenacin biosynthesis and the formation of the avenacin gene cluster, summarised here in two parts: 1) comparative analysis of avenacin biosynthesis amongst species within the *Aveninae* and 2) molecular evolutionary study of triterpene biosynthetic genes in monocots.

7.1.1.Part 1: Survey of avenacin biosynthetic genes amongst species within the *Aveninae*

Chapter 2 describes the dissection of avenacin biosynthesis across the oat (*Avena*) phylogeny both to establish its evolutionary boundary and to provide a foundation upon which to study the formation of the avenacin biosynthetic pathway. Phylogenetic analyses of three molecular markers, *matK*, *TrnL-F* spacer and ITS sequences, revealed that the genus *Avena* fundamentally consists of two different genome types, the A and C genomes, which have diverged from their common ancestor *Avenastrum* early on during species radiation. Screens for avenacin-associated root fluorescence followed by metabolite analysis in ten different oat accessions revealed that avenacin biosynthesis is common to both A and C genome *Avena* species, and is likely to be an ancestral feature of oats; also that it is unique to oats as previously reported (Crombie and Crombie, 1986). Of note, the screens reconfirmed the previous finding that the primitive oat natural variant *A. longiglumis* is avenacin deficient (Osbourn et al., 1994). The C genome oats were found to produce avenacin A-1, B-1 and A-2, but to a lesser extent compared to their A genome counterparts.
Examination of the presence/absence of the five characterised *Sad* genes, which are components of the avenacin biosynthetic pathway, and their expression

profiles is described in Chapter 3. Wheat and *B. distachyon* are also included for a broader examination within the *Pooideae* lineages. Southern blot analysis showed that closely related homologues of *Sad1, 2, 7, 9* and *10* are present in the genomes of all the surveyed oat species in Chapter 2, including *A. longiglumis*, while distantly related homologues of *Sad1, 7* and 9 appeared to be present in the wheat and *B. distachyon* genomes. RT-PCR and northern blot analyses revealed that transcripts of *Sad1, 2* and *7* gene homologues are exclusively detected in roots of the A genome oat seedlings while they are detected in both leaves and roots of the seedlings of C genome oats. Transcripts of *Sad9* homologues were only detected in the roots for all the *Avena* species examined, with the exception of the avenacin-deficient *A. longiglumis*. Furthermore, leaf transcripts of a distantly related *Sad1* homologue were detected in wheat.

Chapter 4 describes the analysis of the *Sad* gene homologue coding sequences, retrieved from genomic DNA and total RNA respectively of different oat species, at the level of sequence similarity and gene product functionality. It was found that the root-expressed transcripts of *Sad* gene homologues of all examined *Avena* species were full-length and likely to be functional, with a high level of sequence similarity within the A and the C genome species but not between the two genome types. Overall, the *Sad* genes are highly conserved amongst different *Avena* species, suggesting a common origin of the avenacin biosynthetic pathway in an ancestral oat lineage. The non-synonymous differences amongst homologues of *Sad1, 2, 7* and *9* from different oat species reflect the independent evolution of these genes after species divergence within the genus *Avena*. Phylogenetic analysis carried out for the sequences of *Sad1, 2* and *7* homologues indicated that the root-expressed *Sad* gene transcripts are likely to have originated from the retrieved genomic sequence of the corresponding A genome species, with the exception of the Ad genome diploid species *A. damascena*. Furthermore, the genomic sequences of *Sad1, 2* and *7* homologues of the C genome oats are the likely donors of the root transcripts, while the leaf transcripts detected in the C genome oats originate from closely related paralogues. These leaf-expressed *Sad* gene paralogues are likely to have arisen after the divergence of the C genome from the A genome oats.

The effects on protein conformation brought about by the non-synonymous differences detected within the predicted amino acid sequences of different oat SAD1 orthologues were further analysed via protein modelling. It was found that SAD1 orthologues of the examined oat species possess highly conserved residues in the active sites, while the non-synonymous differences detected in the sequence analysis are mainly located at the surface of the enzyme, which is more tolerant to mutations (Wagner, 2008).

7.1.2.Part 2: Molecular evolutionary study of the monocot triterpene biosynthetic genes

Phylogenetic analyses were carried out to investigate the evolution of five gene families: oxidosqualene cyclases (OSC), cytochromes P450 family 51 (CYP51), Clade1A serine carboxyl peptidase-like acyltransferases (SCPL), Class I O′-methyl transferases and Group L UDP-glycosyltransferases (UGT). The study of these five genes families, all implicated in triterpene biosynthesis within monocots, is reported in Chapter 5, offering a broader view of how the avenacin biosynthetic pathway might have evolved within the *Poaceae*. Phylogenetic analyses revealed that the triterpene biosynthetic genes largely evolved via tandem duplication, potentially generating the raw genomic material for neo-functionalisation of gene duplicates to encode functionally diverse triterpene biosynthetic enzymes in cereals. Signals of positive selection were detected in events of gene divergence from primary to secondary metabolism, whereas the later gene specialisation events occurred mainly under purifying selection. In addition, radical changes in gene GC content (in *Sad1, 2*, and *7*) and gene structure (in *Sad7*) were observed during the evolution of *Sad* gene homologues. Examination of the genomic distribution of triterpene biosynthetic genes in cereal genomes revealed local enrichment of these genes in regions of *B. distachyon* chromosome 3 and *O. sativa* chromosome 11 that share no syntenic relationship (Schnable et al., 2012), suggesting that colocalisation of triterpene biosynthetic genes may have repeatedly occured in the two plant species.

Syntenic mapping of the gene trees of homologues of *Sad1, 2, 7, 9* and *10* (discussed in Chapter 6) has offered an outline model of the evolution of the avenacin gene cluster (Figure 6.9) that suggests the emergence of precursors of the key avenacin biosynthetic genes *Sad1* and *2* via the ρ-WGD event (70 mya). The key events leading to the emergence of the ancestors of *Sad7* and *Sad9*, occurred subsequently before the divergence of the *Poaceae* and BEP families ($\sim$ 50 mya and $\sim$ 35 mya). The emergence of *Sad1, 2, 7, 9* and *10* for avenacin biosynthesis, however, is found to have occurred recently after species divergence of the subtribe *Aveninae* (<20 mya).

Changes in GC content, especially at the third codon positions (GC_3), during the evolution of the *Sad1* and *2* gene families, were further investigated in Chapter 6. It was found that the GC landscape of *Sad2* homologues was likely to be under selective constraint, while the mechanism of GC change is as yet unidentified.

7.2.The formation of the avenacin gene cluster

The results presented in this thesis collectively show that avenacin biosynthesis has recently evolved within the genus *Avena*. The five proposed evolutionary models of plant metabolic gene cluster formation are outlined here prior to discussion of the evolution of the avenacin gene cluster.

Comparing the thalianol to the marneral gene clusters (Subection 1.4.1) of *A. thaliana* has led to the conclusion that these gene clusters may have arisen from segmental duplication of a physically linked *OSC/CYP705* gene pair (Figure 7.1a) (Field et al., 2011). The two gene clusters then evolved independently via individual recruitment of *THAH* and *MRO* and acquisition of the epigenetic marks for root-specific expression (Field et al., 2011).

The two diterpene gene clusters in rice were proposed to have evolved via repeated recruitment of the CPS1/KSL1 gene pair from gibberellin (GA) biosynthesis (Subsection 1.4.2, Figure 7.1b) to give the CPS2/KSL7 and the CPS4/KSL4 gene pairs respectively (Swaminathan et al., 2009). The two diterpene gene clusters then evolved independently. The phytocassane gene cluster recruited the CYP71z, CYP76M and KSL5/6 genes (Swaminathan et al., 2009) and subsequently expanded by duplication of genes within the gene cluster for the biosynthesis of phytocassanes A to E (Swaminathan et al., 2009). On the other hand, CYP99A and OsMAS were recruited to the momilactone gene cluster (Swaminathan et al., 2009).

Benzoxazinoid biosynthesis evolved in monocot and dicot species via repeated evolution of *Bx1* and *Bx8* (Subsection 1.4.3, Figure 7.1c) (Dick et al., 2012; Schullehner et al., 2008). A phylogenomic study reported that physical linkage of the ancestral '*Bx1*' and '*Bx2*' genes initiated the formation of DIBOA biosynthesis as well as gene cluster formation in the *Poaceae* ancestor (Dutartre et al., 2012). The DIBOA gene cluster is intact in *Z. mays* but is split into two sub-cluster in the wheat and rye genomes and absent within the BEP family members (Dutartre et al., 2012; Sue et al., 2011).

Cyanogenic glycoside biosynthetic genes were found to have evolved repeatedly in three plant species and one arthropod species (Subsection 1.4.4, Figure 7.1d) (Takos et al., 2011). In the three plant species, it is believed that the broad substrate specificity of CYP79 has led to its repeated recruitment to oxime production, the first committed step for cyanogenic glycoside biosynthesis, followed by recruitment of the downstream CYP and UGT genes to the pathway (Takos et al., 2011). Genomic analysis revealed independent assembly of the three plant cyanogenic glycoside gene clusters (Takos et al., 2011).

The functional gene cluster for terpene biosynthesis in tomatoes commenced with

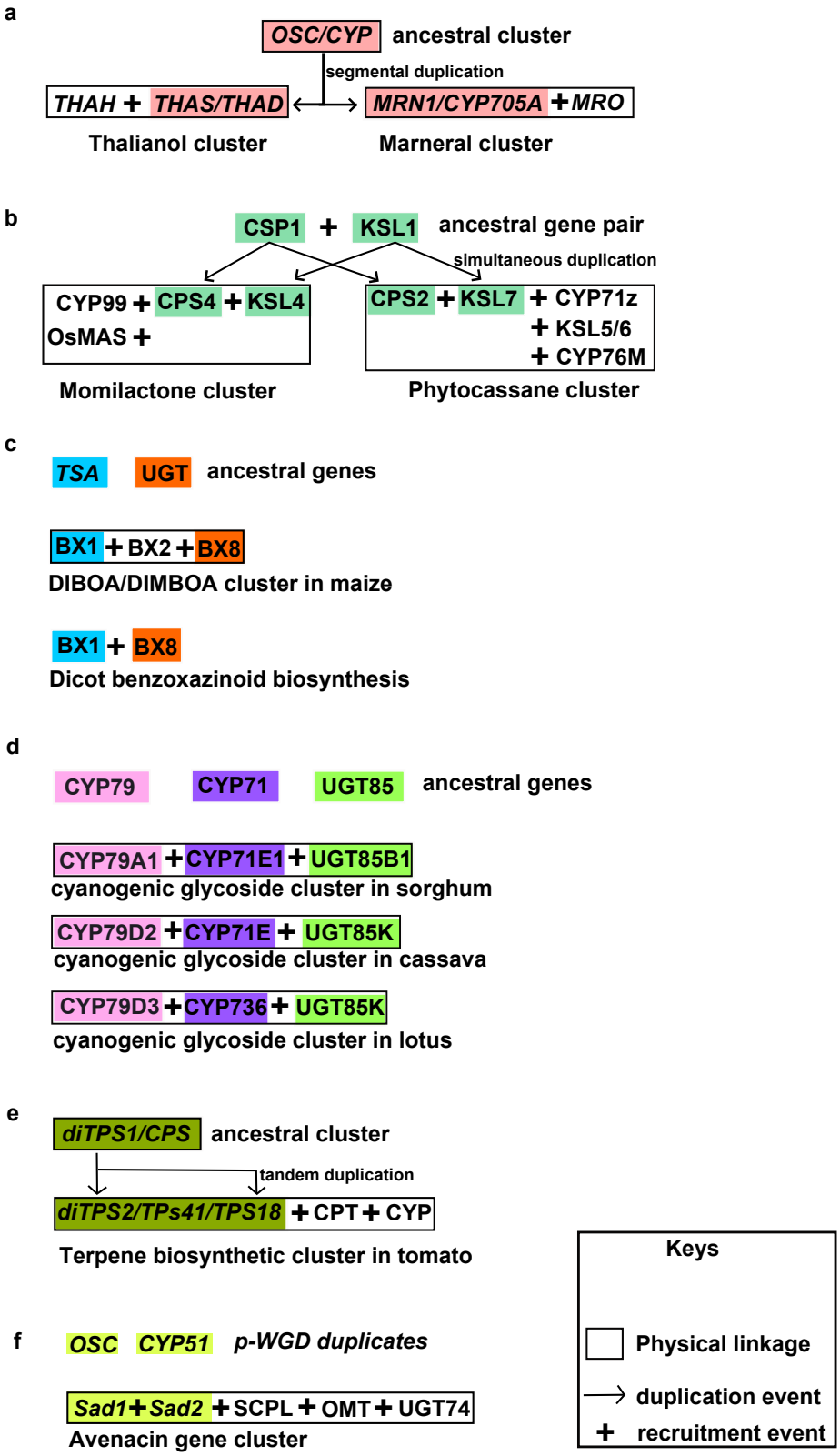


Figure 7.1: The key events proposed to have occured in the evolution of plant gene clusters.

the physical linkage of the diterpene biosynthetic genes *TPS41* (terpene synthase 41) and *diTPS1* (diterpene synthase 1) (Figure 7.1e) (Matsuba et al., 2013). The gene cluster further expanded via tandem duplication and functional divergence of the *diTPS1* genes within the gene cluster, and recruitment of *cis*-prenyl transferase (CPT) and cytochrome P450 genes (Matsuba et al., 2013).
These five evolutionary models indicate that common mechanisms are likely to underlie the assembly of both the plant gene cluster and the pathway they encode. In summary, plant gene clusters are likely to be founded by the formation of the 'signature' genes encoding the branch point enzymes of the specialised metabolic pathway (Figure 7.1). The initial clustering of two or more functionally-related genes with the 'signature' genes appears to contribute to the subsequent pathway elongation and expansion of the gene cluster, leading to the establishment of the present-day clusters.

7.2.1.The likely scenario of avenacin gene cluster formation

The analyses described in this thesis indicate that the ρ-WGD event led to the emergence of precursor *OSC* and *CYP51* sequences for triterpene biosynthesis in monocots (Figure 7.1f). The event of simultaneous recruitment of the *OSC* and *CYP51* genes to triterpene biosynthesis from *CAS1* and *CYP51G1*, which are involved in sterol metabolism, is likely to have marked the starting point of the evolution of avenacin biosynthesis. This ancient *OSC/CYP* gene pair has undergone independent evolution in different monocot lineages and has evolved to synthesise avenacin in oats.
The closest homologues of *A. strigosa Sad1* were found in *S. bicolor* and *Z. mays*, suggesting repeated evolution of β-amyrin biosynthesis, or alternatively the loss of corresponding *OSC* homologues in rice and *B. distachyon*. The lack of closely related *Sad* gene homologues in wheat, as indicated in Chapters 3 and 5, further supports the absence of closely related *Sad1* homologues within members of the BEP family. It also suggests that the avenacin biosynthetic gene *Sad1* must have evolved independently after divergence of the *Aveninae* subtribe from other grasses. If that is the case, subsequent gene recruitment events for avenacin biosynthetic enzymes downstream of SAD1 must also have occurred within the *Aveninae*.
Phylogenetic analysis described in Chapter 5 also indicates that the triterpene biosynthetic genes have evolved largely via tandem duplication throughout monocot evolution. These tandem duplication events followed by functional diversifications of the tandem copies may have provided a toolkit of

multi-functional 'tailoring' enzymes to modify the triterpene scaffolds generated by the *SAD1* homologues. Gene recruitment events from tandem duplicates of *Sad7, 9* and *10* gene family members may have led to the establishment of the complete avenacin biosynthetic pathway (Figure 7.1f).

In the future, surveys on the absence/presence of *Sad* gene homologues within the closely related species of *Avena* such as *Sesleria*, *Helictotrichon* and *Avenula* (Figure 2.1) will enable a comprehensive analysis of the boundary of avenacin biosynthesis evolution. Two BAC libraries of *A. longiglumis* and *A. clauda* have recently been constructed and the sequenced genome of *A. atlantica* will be available in the near future. Investigating the genomic distribution of *Sad* gene homologues using these resources will reveal the avenacin gene cluster organisations of different *Avena* species and will enable further investigations of the evolution of the gene cluster structure following species radiation.

7.3.Future perspective

7.3.1.Bioinformatics driven genome mining for novel metabolic gene clusters

The development of high-throughput DNA sequencing and metabolic profiling technologies are revolutionising biological research, increasing both the depth and breadth of our studies of biological systems. For example, these new technologies are enabling studies of specialised metabolism in plants via a combination of data-driven system biology and functional genomics approaches, opening up the new research field coined 'phytochemical genomics' (Saito, 2013). Utilising the principles of gene clustering is one of the most promising ways to identify new metabolic pathways and to facilitate our navigation through the wealth of omics data. Systematic prediction of microbial secondary metabolic gene clusters using software such as SMURF (Khaldi et al., 2010), antiSMASH (Blin et al., 2013) and MultiGeneBlast (Medema et al., 2013) followed by reverse genetics experiments have greatly accelerated the process identifying gene within novel pathways. These systematic approaches could be also adopted to mining plant genome sequence data for novel gene clusters.

Besides co-localisation of functionally related genes, co-ordinated gene expression is also a feature of operon-like gene clusters. An example of making use of this feature is the 'guilt-by-association' study using both gene clustering information and gene co-expression data to direct mining of cytochrome P450s for triterpene biosynthesis (Castillo et al., 2013). Another feature of gene clustering is the co-inheritance of combinations of beneficial alleles. Applying

this feature of gene clusters to genome-wide association studies (GWAS) amongst natural varieties of plants possessing diverse metabolite profiles is another strategy for the identification of co-localised and co-evolving alleles that are likely to be implicated in the same pathway (Saito, 2013).
With the increasing volume of identified gene clusters, phylogenetic and comparative analyses can be carried out to deconvolute the selective pressures behind gene cluster formation and may enable researchers to examine more closely how environmental adaption inferences the genomic evolution of an organism.

7.3.2. Synthetic engineering of gene clusters complying with rules of nature

Bottom-up engineering

Synthetic biology will provide an excellent platform to generate bottom-up design-based gene clusters in the future, when the components and rules of metabolic gene clusters are well defined (Table 7.1). Taking similar approaches to IGEM (International Genetically Engineered Machine) (Smolke, 2009), standardized biobricks of plant secondary metabolic genes will be easily built and assembled to create synthetic gene clusters (Xu and Koffas, 2013) via the new DNA recombination techniques (reviewed in Ellis et al. (2011)). The well-developed heterologous expression platforms of *Escherichia coli* and *Saccharomyces cerevisiae* and a growing synthetic toolkit for genetic engineering of cyanobacteria will provide a suite of synthetic gene cluster hosts for industrial-scale production of novel metabolites (Ducat et al., 2011; Khalil et al., 2012; Nakagawa et al., 2011; Siddiqui et al., 2012). The pEAQ expression vector (tailored from the Cowpea mosaic virus RNA-2 (CPMV-*HT*)) is now routinely used for transient expression in leaves of *Nicotiana benthamiana* to produce high yield of foreign proteins and has also been shown to be an effective heterologous expression platform for the (re)construction of plant specialised metabolic pathways (Sainsbury et al., 2009). Synthetic gene clusters are also a promising test bed for 'combinatorial biosynthesis', to generate new-to-nature compounds through mix and match of specialised metabolic genes from different pathways (Facchini et al., 2012).

Rules for gene cluster construction

Successful gene cluster design is likely to possess the following properties: 1) the biochemical compatibilities of the biosynthetic parts (whether the enzyme

Synthetic gene cluster component	Rule of good design
Coding sequence of enzymes	Feasible combination of genes
Transcriptional promoters and terminators	Architecture of the gene cluster
Ribosomal binding sites	
Organelle targetting signals	
Markers for chromatin remodelling	

Table 7.1: Summary of designing synthetic gene clusters.

combinations are likely to form a functional pathway), 2) the optimisation of gene expression for correct stoichiometry of gene products and 3) prevention of interference from the synthetic pathway with the cellular metabolic network of the host. Comparative studies of gene clusters and metabolic pathways existing in nature will provide us with guidelines for designing gene clusters, such as biochemically feasible combinations of enzymes, orientation of genes and regulatory elements in the gene cluster, and the optimal location for insertion of the synthetic gene cluster within the host genome. The mechanism of transcriptional regulation of plant gene clusters at the chromatin level (Field et al., 2011; Krokida et al., 2013) is also a critical area to exploit, especially for synthetic gene clusters designed for a multi-cellular host, to confer tissue specific expression for efficient harvesting of desired products and insulation from antagonistic pathways.

7.3.3.Concluding remarks

The investigation of the evolution of avenacin biosynthesis and gene cluster formation has increased our current understanding of how the chemodiversity of triterpene biosynthesis has been achieved in monocots. Furthermore, an evolutionary framework for the formation of the avenacin biosynthetic pathway and gene cluster has been formulated. The growing number of discovered of plant metabolic gene clusters will provide valuable data for further investigations into how the interplay of genomic plasticity and environmental adaption has led to metabolic innovation and gene clustering in plants.

Appendix

1.1.List of supplimentary data

The supplimentary data for Chapter 2, 3, and 4 were stored in a DVD. There are multiple formats of files and required UCSF Chimera 1.6.2 (Pettersen et al., 2004), Microsoft Excel 2010, MEGA 5.1 (Tamura et al., 2011) and FigTree v 1.3.1 (available at http://tree.bio.ed.ac.uk/software/figtree/).

Chapter	File directory	File name	description
Chapter 2	Chapter 2/oat phylogeny study	ITS_MAtK_TRNL _table.xlsx	Table 2.1 Summary of ITS, matK and TrnL-F sequences of the oat species.
	Chapter 2/	combined LCMS data.xlsx	Table 2.2 Raw data from five set of LCMS experiments.
	Chapter 2/	UV screen photographs of roots of seedlings.docx	Photographic records of root fluorescence.
	Chapter 2/oat phylogeny study	Tree 2.1 TrnL_F_RAxML_ name.mts	Phylogenetic trees TrnL-F sequence alignments in RAxML7.0.3.
	Chapter 2/oat phylogeny study	Tree 2.2 TrnL_F_MB_name.mts	Phylogenetic trees of TrnL-F sequence alignments in MrBayes 3.2.1.
	Chapter 2/oat phylogeny study	TrnL_F_refin_align.phy	Alignment of TrnL-F sequences.
	Chapter 2/oat phylogeny study	Tree 2.3 MatK_RAxML_name.mts	Phylogenetic trees matK sequence alignments in RAxML7.0.3.
	Chapter 2/oat phylogeny study	Tree 2.4 MatK_MB_name.mts	Phylogenetic trees of matK sequence alignments in MrBayes 3.2.1.
	Chapter 2/oat phylogeny study	MatK_align_refine.phy	Alignment of matK sequences.
	Chapter 2/oat phylogeny study	Tree 2.5 ITSp_RAxML_name.mts	Phylogenetic trees of ITS1-5.8r-ITS2 sequence alignments in RAxML7.0.3 with partitioning.
	Chapter 2/oat phylogeny study	Tree 2.6 ITSp_MB_name.mts	Phylogenetic trees of ITS1-5.8r-ITS2 sequence alignments in MrBayes 3.2.1 with partitioning.
	Chapter 2/oat phylogeny study	ITS_rc_alignp.phy	Alignment of ITS1-5.8r-ITS2 sequences
	Chapter 2/oat phylogeny study	Tree 2.7 Removal of duplicates RAxML Supertree.tre	Supertree constructed using the RAXML source trees of matK, TrnL-F and ITS sequences with duplicated species tips removed

continued on next page

continued from previous page

	Chapter 2/oat phylogeny study	Tree 2.8 Removal of duplicates MB Supertree.tre	Supertree constructed using the MrBayes source trees of matK, TrnL-F and ITS sequences with duplicated species tips removed
	Chapter 2/oat phylogeny study	Tree 2.9 Removal of polyploid RAxML Supertree.tre	Supertree constructed using the RAXML source trees of matK, TrnL-F and ITS sequences with duplicated and polyploid species tips removed
	Chapter 2/oat phylogeny study	Tree 2.10 Removal of polyploid MB Supertree.tre	Supertree constructed using the mrBayes source trees of matK, TrnL-F and ITS sequences with duplicated and polyploid species tips removed
	Chapter 2/oat phylogeny study	Tree 2.11 concat_nonpart_Mb.mts	Phylogenetic trees of concatenated ITS, matK and TrnL-F sequence alignments in MrBayes 3.2.1 with partitioning.
	Chapter 2/oat phylogeny study	Tree 2.12 concat_nonpart_RAxML.mts	Phylogenetic trees of concatenated ITS, matK and TrnL-F sequence alignments in RAxML 7.0.3 with partitioning.
	Chapter 2/oat phylogeny study	Tree 2.13 concat_part_Mb.mts	Phylogenetic trees of concatenated ITS, matK and TrnL-F sequence alignments in MrBayes 3.2.1 without partitioning.
	Chapter 2/oat phylogeny study	Tree 2.14 concat_part_RAxML.mts	Phylogenetic trees of concatenated ITS, matK and TrnL-F sequence alignments in RAxML 7.0.3 without partitioning.
	Chapter 2/oat phylogeny study	Concaternate coregroup.fas	Concaternated alignment of ITS, matK and TrnL-F sequences.

continued on next page

continued from previous page

Chapter 3	Chapter 3	Loading of Xba1 DNA digests.pdf	Ethidium bromide-stained gel image showing the DNA loading of the *Xba1* DNA digests of oat species.
Chapter 4	Chapter 4/RT-PCR transcript lignment	Table 4.1 sequencecomparison.xlsx	non-synonymous differences amongst oat transcripts
	Chapter 4/RT-PCR transcript lignment	Alignment 4.1 Sad1_codon_aligned.fas	codon alignment of Sad1 trnascripts
	Chapter 4/RT-PCR transcript lignment	Alignment 4.2 Sad2_codon_aligned.fas	codon alignment of Sad1 trnascripts
	Chapter 4/RT-PCR transcript lignment	Alignment 4.3 Sad7_codon_aligned.fas	codon alignment of Sad1 trnascripts
	Chapter 4/RT-PCR transcript lignment	Alignment 4.4 Sad9_codon_aligned.fas	codon alignment of Sad1 trnascripts
	Chapter 4/Sequencing analysis of Sad genes/Sad1	Tree 4.1 RAxML_Sad1aa.tre	Phylogenetic tree of Sad1 amino acid alignment generated in RAxML 7.0.3.
	Chapter 4/Sequencing analysis of Sad genes/Sad1	Tree 4.2 RAxML_S1cds.tre	Phylogenetic tree of Sad1 codon alignment generated in RAxML 7.0.3.
	Chapter 4/Sequencing analysis of Sad genes/Sad1	Tree 4.3 MB_S1aa.tre	Phylogenetic tree of Sad1 amino acid alignment generated in MrBayes 3.2.1.
	Chapter 4/Sequencing analysis of Sad genes/Sad1	Tree 4.4 MB_S1cds.tre	Phylogenetic tree of Sad1 codon alignment generated in MrBayes 3.2.1.
	Chapter 4/Sequencing analysis of Sad genes/Sad1	S1aa_RT_ gs_alnREFINE.phy	Amino acid alignment of Sad1 homologues.
	Chapter 4/Sequencing analysis of Sad genes/Sad1	S1CDS_RT_ gs_alnREFINE.phy	Codon sequence alignment of Sad1 homologues.
	Chapter 4/Sequencing analysis of Sad genes/Sad2	Tree 4.5 RAxML_S2_aa.tre	Phylogenetic tree of Sad2 amino acid alignment generated in RAxML 7.0.3.

continued on next page

continued from previous page

	Chapter 4/Sequencing analysis of Sad genes/Sad2	Tree 4.6 RAxML_S2_cds.tre	Phylogenetic tree of Sad2 codon alignment generated in RAxML 7.0.3.
	Chapter 4/Sequencing analysis of Sad genes/Sad2	Tree 4.7 MB_S2_aa.tre	Phylogenetic tree of Sad2 amino acid alignment generated in MrBayes 3.2.1.
	Chapter 4/Sequencing analysis of Sad genes/Sad2	Tree 4.8 MB_S2_cds.tre	Phylogenetic tree of Sad2 codon alignment generated in MrBayes 3.2.1.
	Chapter 4/Sequencing analysis of Sad genes/Sad2	S2_RT_gs_ref_aa_aln_refined.phy	Amino acid alignment of Sad2 homologues.
	Chapter 4/Sequencing analysis of Sad genes/Sad2	S2_RT_gs_ref_codon_aln_refined.phy	Codon sequence alignment of Sad2 homologues.
	Chapter 4/Sequencing analysis of Sad genes/Sad7	Tree 4.9 RAxML_Sad7cds_gsRTref.tre	Phylogenetic tree of Sad7 amino acid alignment generated in RAxML 7.0.3
	Chapter 4/Sequencing analysis of Sad genes/Sad7	Tree 4.10 RAxML_Sad7_aa_gsRTref.tre	Phylogenetic tree of Sad7 codon alignment generated in RAxML 7.0.3
	Chapter 4/Sequencing analysis of Sad genes/Sad7	Tree 4.11 MB_Sad7_aa_gsRTref.tre	Phylogenetic tree of Sad7 amino acid alignment generated in MrBayes 3.2.1
	Chapter 4/Sequencing analysis of Sad genes/Sad7	Tree 4.12 MB_Sad7_cds_gsRTref.tre	Phylogenetic tree of Sad7 codon alignment generated in MrBayes 3.2.1
	Chapter 4/Sequencing analysis of Sad genes/Sad7	S7_gs_ref_RT_aa_aln_ref.phy	Amino acid alignment of Sad7 homologues.
	Chapter 4/Sequencing analysis of Sad genes/Sad7	S7_gs_ref_RT_codon_aln_ref.phy	Codon sequence alignment of Sad7 homologues.
	Chapter 4/Sad1 protein models	1w6k_Asbas.pyc	Structural alignment of 1w6k and AsBas.
	Chapter 4/Sad1 protein models	All aligned.py	Structural alignment of 1w6k and all protein models of Sad1 homologues.
	Chapter 4/Sad1 protein models	All aligned.pyc	Structural alignment of 1w6k and all protein models of Sad1 homologues.

continued on next page

continued from previous page

	Chapter 4/Sad1 protein models	Alignment 4.5 Sad1_structural alignment.afasta	Structural alignment all protein models of Sad1 homologues.
	Chapter 4/Sad1 protein models	Table 4.2 ITASSER summary.xlsx	Details in protein models of Sad1 homologues generated in ITTASSER 2.0.1 server.
Chapter5	Chapter 5/Sad7 analysis/HMMER	HMMprofileSad7.hmm	HMMprofile bult from amino acid alignment of Sad7 closest homologues.
	Chapter 5/Sad7 analysis/HMMER /e-50NJtree	Sad7_aligned _stripcoloum.fas	Amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad7 analysis/HMMER /e-50NJtree	Tree 5.1 minimaltree.mts	BIONJ tree built from the minimal amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad7 analysis/HMMER /e-50NJtree	Tree 5.2 stripcoltree.mts	BIONJ tree built from the strip_column amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad7 analysis/HMMER /e-50NJtree	Tree 5.3 - 50NJ_SCPL.mts	BIONJ tree of monocot SCPL built from the minimal amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad7 analysis/SCPL analysis	Tree 5.4 SCPL_aa_RAxML.mts	Phylogenetic tree of SCPL generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad7 analysis/SCPL analysis	Tree 5.5 SCPL_cds_RAxML.mts	Phylogenetic tree of SCPL generated from codon alignment in RAxML 7.0.4.
	Chapter 5/Sad7 analysis/SCPL analysis	Tree 5.6 SCPL_aa_MB.mts	Phylogenetic tree of SCPL generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad7 analysis/SCPL analysis	Tree 5.7 SCPL_cds_MB.mts	Phylogenetic tree of SCPL generated from codon alignment in Mabayes 3.2.1.

continued on next page

continued from previous page

	Chapter 5/Sad7 analysis/SCPL analysis	SCPL_aa_align _refined.phy	Amino acid alignment of SCPL genes.
	Chapter 5/Sad7 analysis/SCPL analysis	SCPL_cds_align _refined.phy	Codon alignment of SCPL genes.
	Chapter 5/Sad7 analysis/SCPL analysis	Table 5.1 Gene structure SCPL.xlsx	Summary of gene structure of SCPL Clade 1A homologues.
	Chapter 5/Sad7 analysis/SCPL analysis	Table 5.2 SCPL transcript table.docx	Annotated amino acid and coding sequences of SCPL Clade 1A homologues.
	Chapter 5/Sad7 analysis/Sad7 smaller tree analysis	Tree 5.8 SCPL_aa_RAxML.mts	Phylogenetic tree of Sad7 generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad7 analysis/Sad7 smaller tree analysis	Tree 5.9 SCPL_aa_MB.mts	Phylogenetic tree of Sad7 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad7 analysis/Sad7 smaller tree analysis	Sad7_aa_aligned_ refined.phy	Amino acid alignment of Sad7 genes.
	Chapter 5/Sad7 analysis/Sad7 slection test	Sad7_codon_aln.fas	Codon alignemnt of Sad7 homologues.
	Chapter 5/Sad7 analysis/Sad7 slection test	Table 5.3 Sad7 PAML selection test.xlsx	Summary of PAML branch-site tests of Sad7 homologues.
	Chapter 5/Sad7 analysis/Sad7 slection test	Table 5.4 Sad7 smallertree GCcontent.xlsx	GC content of coding sequences of Sad7 homologues.
	Chapter 5/Sad7 analysis/Sad7 slection test	Sad7_pairewise.txt	PAML pairwise test on Sad7 homologues.
	Chapter 5/Sad9 analysis/HMMER	HMMprofileSad9.hmm	HMMprofile bult from amino acid alignment of Sad9 closest homologues.
	Chapter 5/Sad9 analysis/HMMER /Sad9 e-40NJtree	Sad9HMM_aa_aligned _stripg.fas	Amino acid alignment of e^{-40} hits from HMMsearch.

continued on next page

continued from previous page

	Chapter 5/Sad9 analysis/HMMER /Sad9 e-40NJtree	Sad9HMMaa_aligned _minimal.mts	Tree 5.10 BIONJ tree built from amino acid alignment of e^{-40} hits from HMMsearch.
	Chapter 5/Sad9 analysis/OMT1 analysis	Tree 5.11 OMT_aa_RAxML.mts	Phylogenetic tree of OMT1 generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad9 analysis/OMT1 analysis	Tree 5.12 OMT_cds_RAxML.mts	Phylogenetic tree of OMT1 generated from codon alignment in RAxML 7.0.4.
	Chapter 5/Sad9 analysis/OMT1 analysis	Tree 5.13 OMT_aa_MB.mts	Phylogenetic tree of OMT1 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad9 analysis/OMT1 analysis	Tree 5.14 OMT_cds_MB.mts	Phylogenetic tree of OMT1 generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad9 analysis/OMT1 analysis	OMT_aa_alignrefine.phy	Amino acid alignment of OMT1 genes.
	Chapter 5/Sad9 analysis/OMT1 analysis	OMT_codon_ alignrefine.phy	Codon alignment of OMT1 genes.
	Chapter 5/Sad9 analysis/OMT1 analysis	Table 5.5 OMT_genestructure.xlsx	Summary of gene structure of OMT1 homologues.
	Chapter 5/Sad9 analysis/OMT1 analysis	Table 5.6 OMT transcript table.docx	Annotated amino acid and coding sequences of OMT homologues.
	Chapter 5/Sad9 analysis/Sad9 analysis	Tree 5.15 Sad9_aa_RAxML.mts	Phylogenetic tree of Sad9 generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad9 analysis/Sad9 analysis	Tree 5.16 Sad9_cds_RAxML.mts	Phylogenetic tree of Sad9 generated from codon alignment in RAxML 7.0.4.

continued on next page

continued from previous page

	Chapter 5/Sad9 analysis/Sad9 analysis	Tree 5.17 Sad9_aa_MB.mts	Phylogenetic tree of Sad9 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad9 analysis/Sad9 analysis	Tree 5.18 Sad9_cds_MB.mts	Phylogenetic tree of Sad9 generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad9 analysis/Sad9 analysis	Sad9_aa_alignrefine.phy	Amino acid alignment of Sad9 genes.
	Chapter 5/Sad9 analysis/Sad9 analysis	Sad9_codon_alignrefine .phy	Codon alignment of Sad9 genes.
	Chapter 5/Sad9 analysis/Sad9 slection test	Table 5.7 Sad9 selection summary table.xlsx	Summary of PAML branch-site tests of Sad9 homologues.
	Chapter 5/Sad9 analysis/Sad9 slection test	Table 5.8 GC content of Sad9.xlsx	GC content of coding sequences of Sad9 homologues.
	Chapter 5/Sad9 analysis/Sad9 slection test	Sad9_pairwisedNdS.txt	PAML pairwise test on Sad9 homologues.
	Chapter 5/Sad10 analysis/HMMER	HMMprofileSad10 csh.hmm	HMMprofile bult from amino acid alignment of Sad10 closest homologues.
	Chapter 5/Sad10 analysis/HMMER /e-50NJtree	Sad10_nj_50align_edit.fas	Amino acid alignment of e^{-50} hits from HMM-search.
	Chapter 5/Sad10 analysis/HMMER /e-50NJtree	Tree 5.19 Sad10_nj_50align.mts	BIONJ tree built from amino acid alignment of e^{-50} hits from HMM-search.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.20 LGT_aa_RAxML.mts	Phylogenetic tree of GroupL GT generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.21 LGT_cds_RAxML.mts	Phylogenetic tree of GroupL GT generated from codon alignment in RAxML 7.0.4.

continued on next page

continued from previous page

	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.22 LGT_aa_MB.mts	Phylogenetic tree of GroupL GT generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.23 LGT_cds_MB.mts	Phylogenetic tree of GroupL GT generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.24 LGT_cds _3rd_M_B_MB.mts	Phylogenetic tree of GroupL GT generated from codon alignment in Mabayes 3.2.1 without thrid codon positions.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Tree 5.25 LGT_cds _3bs_RAxML.mts	Phylogenetic tree of GroupL GT generated from codon alignment in RAxML 7.0.4 without thrid codon positions.
	Chapter 5/Sad10 analysis/GroupL GT analysis	LGT_aa_alignrefine.phy	Amino acid alignment of GroupL GT genes.
	Chapter 5/Sad10 analysis/GroupL GT analysis	LGT_cds_alignrefine.phy	Codon alignment of GroupL GT genes.
	Chapter 5/Sad10 analysis/GroupL GT analysis	LGT_cds_3rd_b_strip.phy	Codon alignment of GroupL GT genes without the third codon positions.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Table 5.9 GroupLGT_ genestructure.xlsx	Summary of gene structure of GroupL GT homologues.
	Chapter 5/Sad10 analysis/GroupL GT analysis	Table 5.10 LGT transcript table.docx	Annotated amino acid and coding sequences of GroupL GT homologues.
	Chapter 5/Sad10 analysis/Sad10 analysis	Tree 5.26 Sad10_aa_RAxML.mts	Phylogenetic tree of Sad10 generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad10 analysis/Sad10 analysis	Tree 5.27 Sad10_cds_RAxML.mts	Phylogenetic tree of Sad10 generated from codon alignment in RAxML 7.0.4.

continued on next page

continued from previous page

	Chapter 5/Sad10 analysis/Sad10 analysis	Tree 5.28 Sad10_aa_MB.mts	Phylogenetic tree of Sad10 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad10 analysis/Sad10 analysis	Tree 5.29 Sad10_cds_MB.mts	Phylogenetic tree of Sad10 generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad10 analysis/Sad10 analysis	Sad10_aa_alignrefine.phy	Amino acid alignment of Sad10 homologues.
	Chapter 5/Sad10 analysis/Sad10 analysis	Sad10_cds_alignrefine.phy	Codon alignment of Sad10 homologues.
	Chapter 5/Sad10 analysis/Sad10 slection test	Table 5.11 Sad10 selection test.xlsx	Summary of PAML branch-site tests of Sad10 homologues.
	Chapter 5/Sad10 analysis/Sad10 slection test	Table 5.12 GCtable-GroupL glycosyltransferases.xlsx	GC content of coding sequences of GroupL UGTs.
	Chapter 5/Sad10 analysis/Sad10 slection test	Sad10_pairwisedNdS.txt	PAML pairwise test on Sad9 homologues.
	Chapter 5/Sad2 analysis/HMMER	HMMprofileCyp51.hmm	HMMprofile bult from amino acid alignment of Sad2 closest homologues.
	Chapter 5/Sad2 analysis/HMMER /e-50NJtree	50_nj_aa _aln_refin.fas	Amino acid alignment of e^{-50} hits from HMM-search.
	Chapter 5/Sad2 analysis/HMMER /e-50NJtree	Tree 5.30 Sad2_50nj_aa_aln.mts	BIONJ tree built from amino acid alignment of e^{-50} hits from HMM-search.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 tree no CYPH5 and G4	Tree 5.31 CYP51_aa_RAxML.mts	Phylogenetic tree of CYP51 generated from amino acid alignment without CYP51H5 and CYP51G4 in RAxML 7.0.4.

continued on next page

continued from previous page

	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 tree no CYPH5 and G4	Tree 5.32 CYP51_cds_RAxML.mts	Phylogenetic tree of CYP51 generated from codon alignment without CYP51H5 and CYP51G4 in RAxML 7.0.4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 tree no CYPH5 and G4	Tree 5.33 CYP51_aa_MB.mts	Phylogenetic tree of CYP51 generated from amino acid alignment without CYP51H5 and CYP51G4 in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 tree no CYPH5 and G4	Tree 5.34 CYP51_cds_MB.mts	Phylogenetic tree of CYP51 generated from codon alignment without CYP51H5 and CYP51G4 in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	Tree 5.35 CYP51_aa_RAxML.mts	Phylogenetic tree of CYP51 generated from amino acid alignment containing CYP51H5 and CYP51G4 in RAxML 7.0.4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	Tree 5.36 CYP51_cds_RAxML.mts	Phylogenetic tree of CYP51 generated from codon alignment containing CYP51H5 and CYP51G4 in RAxML 7.0.4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	Tree 5.37 CYP51_aa_MB.mts	Phylogenetic tree of CYP51 generated from amino acid alignment containing CYP51H5 and CYP51G4 in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	Tre 5.38 CYP51_cds_MB.mts	Phylogenetic tree of CYP51 generated from codon alignment containing CYP51H5 and CYP51G4 in Mabayes 3.2.1.

continued on next page

continued from previous page

	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	CYP51_aa_aln_refine2.phy	Amino acid alignment of CYP51 genes containing CYP51H5 and CYP51G4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51 trees contain OSCH5 and G4	CYP51_codon_aln_refine2.phy	Codon alignment of CYP51 genes containing CYP51H5 and CYP51G4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51_3bs	Tree 5.39 Sad2_cds3bs_RAxML.mts	Phylogenetic tree of CYP51 generated from codon alignment without third codon positions in RAxML 7.0.4.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51_3bs	Tree 5.40 Sad2_cds3bs_MB.mts	Phylogenetic tree of CYP51 generated from codon alignment without third codon positions in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/CYP51 analysis/CYP51_3bs	CYP51_3bs.phy	Codon alignment of CYP51s without third codon positions.
	Chapter 5/Sad2 analysis/CYP51 analysis	CYP51_aa_aln_refine.phy	Amino acid alignment of CYP51s genes without CYP51H5 and CYP51G4.
	Chapter 5/Sad2 analysis/CYP51 analysis	CYP51_codon_aln_refine.phy	Codon alignment of CYP51s genes without CYP51H5 and CYP51G4.
	Chapter 5/Sad2 analysis/CYP51 analysis	Table 5.13 Sad2_genestructure.xlsx	Summary of gene structure of CYP51 homologues.
	Chapter 5/Sad2 analysis/CYP51 analysis	Table 5.14 Sad2_transcript_table.docx	Annotated amino acid and coding sequences of CYP51s.
	Chapter 5/Sad2 analysis/Sad2 analysis	Tree 5.41 Sad2_aa_RAxML.mts	Phylogenetic tree of Sad2 generated from amino acid alignment in RAxML7.0.4
	Chapter 5/Sad2 analysis/Sad2 analysis	Tree 5.42 Sad2_cds_RAxML.mts	Phylogenetic tree of Sad2 generated from codon alignment in RAxML7.0.4

continued on next page

continued from previous page

	Chapter 5/Sad2 analysis/Sad2 analysis	Tree 5.43 Sad2_aa_MB.mts	Phylogenetic tree of Sad2 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/Sad2 analysis	Tree 5.44 Sad2_cds_MB.mts	Phylogenetic tree of Sad2 generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad2 analysis/Sad2 analysis	Sad2_aa_alignrefine.phy	Amino acid alignment of Sad2 homologuees.
	Chapter 5/Sad2 analysis/Sad2 analysis	Sad2_cds_alignrefine.phy	Codon alignment of Sad2 homologues.
	Chapter 5/Sad2 analysis/Sad2 selection test	Table 5.15 CYP51_GCtable.xlsx	GC content of coding sequences of Sad2 homologues.
	Chapter 5/Sad1 analysis/Sad2 analysis	Sad2_pairwisedNdS.txt	PAML pairwise analysis of Sad2 homologues.
	Chapter 5/Sad2 analysis/Sad2 selection test	Table 5.16 Sad2 selection test summary.xlsx	Summary of PAML branch-site tests of Sad2 homologues.
	Chapter 5/Sad1 analysis/HMMER	HMMprofileSad1_ch.hmm	HMMprofile bult from amino acid alignment of Sad1 closest homologues.
	Chapter 5/Sad1 analysis/HMMER /e-50NJtree	NJ_50_S1_aa_aln_refin.fas	Amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad1 analysis/HMMER /e-50NJtree	Tree 5.45 Sad1_aa_50NJ.mts	BIONJ tree built from amino acid alignment of e^{-50} hits from HMMsearch.
	Chapter 5/Sad1 analysis/OSC analysis	Tree 5.46 OSC_aa_RAxML.mts	Phylogenetic tree of OSC generated from amino acid alignment in RAxML 7.0.4.
	Chapter 5/Sad1 analysis/OSC analysis	Tree 5.47 OSC_cds_RAxML.mts	Phylogenetic tree of OSC generated from codon alignment in RAxML 7.0.4.

continued on next page

continued from previous page

	Chapter 5/Sad1 analysis/OSC analysis	Tree 5.48 OSC_aa_MB.mts	Phylogenetic tree of OSC generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad1 analysis/OSC analysis	Tree 5.49 OSC_cds_MB.mts	Phylogenetic tree of OSC generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad1 analysis/OSC analysis	OSC_aa_align _refined.phy	Amino acid alignment of OSC genes.
	Chapter 5/Sad1 analysis/OSC analysis	OSC_codon_align _refined.phy	Codon alignment of OSC genes.
	Chapter 5/Sad1 analysis/OSC analysis	Table 5.16 OSC_gene_structure.xlsx	Gene structure summary of OSC homologues.
	Chapter 5/Sad1 analysis/OSC analysis	Table 5.17 OSC_transcript_ table.docx	Annotated coding sequences and amino acid sequences of OSC homologues.
	Chapter 5/Sad1 analysis/Sad1 analysis	Tree 5.50 Sad1_aa_RAxML.mts	Phylogenetic tree of Sad1 generated from amino acid alignment in RAxML7.0.4.
	Chapter 5/Sad1 analysis/Sad1 analysis	Tree 5.51 Sad1_cds_RAxML.mts	Phylogenetic tree of Sad1 generated from codon alignment in RAxML7.0.4.
	Chapter 5/Sad1 analysis/Sad1 analysis	Tree 5.52 Sad1_aa_MB.mts	Phylogenetic tree of Sad1 generated from amino acid alignment in Mabayes 3.2.1.
	Chapter 5/Sad1 analysis/Sad1 analysis	Tree 5.53 Sad1_cds_MB.mts	Phylogenetic tree of Sad1 generated from codon alignment in Mabayes 3.2.1.
	Chapter 5/Sad1 analysis/Sad1 analysis	Sad1_aa_align.phy	Amino acid alignment of Sad1 homologuees.
	Chapter 5/Sad1 analysis/Sad1 analysis	Sad1_codon_align.phy	Codon alignment of Sad1 homologuees.

continued on next page

continued from previous page			
	Chapter 5/Sad1 analysis/Selection test	Table 5.18 Sad1 PAML test.xlsx	Summary of PAML branch-site test of Sad1 homologues.
	Chapter 5/Sad1 analysis/Selection test	Table 5.19 Sad1_GCtest.xlsx	GC content of coding sequences of Sad1 homologues.
	Chapter 5/Sad1 analysis/Selection test	Sad1_pairwisedNdS.txt	PAML pairwise analysis of Sad1 homologues.
Chapter 6	Chapter 6/	Table 6.1 GC-oat Sad1.xlsx	GC content of exon 1-18 and intron 1-17 of Sad1 homologues.
	Chapter 6/	Figure 6.1 Sad1 GC landscape	GC plot along Sad1 homologues
	Chapter 6/	Table 6.2 GC-oat Sad2. xlsx	GC content of exon 1-2 and intron 1 of Sad2 homologues
	Chapter 6/	Figure 6.2 Sad2 GC landscape	GC plot along Sad2 homologues
	Chapter 6/	Table 6.3 Rice Sad2. xlsx	GC content of exon 1-2 and intron 1 of rice CYP51s

Table 1.1: List of supplimentary data

Supplier details

1.1.List of suppliers

Suppliers for chemical reagents, experimental kits and materials are mentioned by abbreviation the name of the company in the materials and method sections in chapter two, three, and four. The details of each company are listed in Table 1.1.1. Service provider are mentioned by company name in the main text. The deatils of each company are listed in Table 1.2.1

New England Biolabs	NEB	www.neb.com
MERCK United Kingdom	MERCK®	www.merck.co.uk
Qiagen	Qiagen	http://www.qiagen.com
Life Technologies Ltd - Invitrogen	Invitrogen	http://www.lifetechnologies.com /uk/en/home.html
Sigma-Aldrich®	Sigma	http://www.sigmaaldrich.com /united-kingdom.html
GE Healthcare Life Science	GE Healthcare	http://www.gelifesciences.com/ webapp/wcs/stores/servlet /Home/zh/GELifeSciences-UK/
Thermo Fisher Scientific Inc	Thermo®	http://www.fisher.co.uk/

Table 1.1.1: List of suppliers

1.2.List of service providers

Company/Corporate Name	Service provided	Website
JIC Metabolite Service	Liquid chromatography/Mass spectrometry	http://www.jic.ac.uk/ services/metabolomics/index.htm
Genome Enterprise Ltd	DNA sequencing	http://www.genome-enterprise.com/
Eurofins Mwg Operon	DNA sequencing	http://www.eurofinsgenomics.eu/

Table 1.2.1: Details of service providers

Bibliography

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104 – 2105.

Aceituno, F., Moseyko, N., Rhee, S., and Gutiérrez, R. (2008). The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genomics*, 9(1):438.

Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A., and Dopazo, J. (2010). Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS Computational Biology*, 6(10):e1000953.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*, 1(5):543 – 556.

Amoutzias, G. and Van de Peer, Y. (2008). Together we stand: genes cluster to coordinate regulation. *Developmental Cell*, 14(5):640 – 642.

Augustin, J., Kuzina, V., Andersen, S., and Bak, S. (2011). Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry*, 72(6):435 – 457.

Badaeva, E. D., Shelukhina, O. Y., Diederichsen, A., Loskutov, I. G., and Pukhalskiy, V. A. (2010a). Comparative cytogenetic analysis of *Avena macrostachya* and diploid C-genome *Avena* species. *Genome*, 53(2):125 – 137.

Badaeva, E. D., Shelukhina, O. Y., Diederichsen, A., Loskutov, I. G., and Pukhalskiy, V. A. (2010b). Phylogenetic Relationships of Tetraploid AB-Genome *Avena* Species Evaluated by Means of Cytogenetic (C-Banding and FISH) and RAPD Analyses. *Journal of Botany*, 2010.

Bak, S., Paquette, S. M., Morant, M., Morant, A. V., Saito, S., Bjarnholt, N., Zagrobelny, M., Jørgensen, K., Osmani, S., Simonsen, H. T., et al. (2006). Cyanogenic glycosides: a case study for evolution and application of cytochromes P450. *Phytochemistry Reviews*, 5(2-3):309 – 329.

Ballouz, S., Francis, A. R., Lan, R., and Tanaka, M. M. (2010). Conditions for the evolution of gene clusters in bacterial genomes. *PLoS Computational Biology*, 6(2):e1000672.

Barakat, A., Choi, A., Yassin, N. B. M., Park, J. S., Sun, Z., and Carlson, J. E. (2011). Comparative genomics and evolutionary analyses of the O-methyltransferase gene family in Populus. *Gene*, 479(1-2):37 – 46.

Batada, N. N. and Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics*, 39(8):945 – 949.

Bateman, G., Gutteridge, R., Jenkyn, J., Spink, J., and McVittie, J. (2006). Take-all in winter wheat - management guidelines, home grown cereals authority, london, uk.

Baum, B. R. (1977). *Oats: wild and cultivated. A monograph of the genus Avena L.(Poaceae).* Minister of Supply and Services.

Bellamine, A., Lepesheva, G. I., and Waterman, M. R. (2004). Fluconazole binding and sterol demethylation in three CYP51 isoforms indicate differences in active site topology. *Journal of Lipid Research*, 45(11):2000 – 2007.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., Estep, M., Feng, L., Vaughn, J. N., Grimwood, J., et al. (2012). Reference genome sequence of the model plant *setaria*. *Nature biotechnology*, 30(6):555 – 561.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends in Ecology & Evolution*, 19(6):315 – 322.

Birchler, J. A. and Veitia, R. A. (2012). Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746 – 14753.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Research*, 14(5):988 – 995.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0 - a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research*, 41(W1):W204 – W212.

Bornberg-Bauer, E. and Albà, M. (2013). Dynamics and adaptive benefits of modular protein evolution. *Current opinion in structural biology*, pages 459 – 466.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., Bank, M. v. d., Chase, M. W., and Hodkinson, T. R. (2008). Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. *Molecular phylogenetics and evolution*, 47(2):488 – 505.

Bratlie, M. S., Johansen, J., and Drabløs, F. (2010). Relationship between operon preference and functional properties of persistent genes in bacterial genomes. *BMC Genomics*, 11(1):71.

Brendel, V., Xing, L., and Zhu, W. (2004). Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, 20(7):1157 – 1169.

Brown, C. A., Murray, A. W., and Verstrepen, K. J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology*, 20(10):895 – 903.

Buetti-Dinh, A., Ungricht, R., Kelemen, J. Z., Shetty, C., Ratna, P., and Becskei, A. (2009). Control and signal processing by transcriptional interference. *Molecular Systems Biology*, 5(300).

Burkhardt, H. J., Maizel, J. V., and Mitchell, H. K. (1964). Avenacin, an Antimicrobial Substance Isolated from Avena sativa. II. Structure*. *Biochemistry*, 3(3):426 – 431.

Busk, P. K. and Møller, B. L. (2002). Dhurrin synthesis in sorghum is regulated at the transcriptional level and induced by nitrogen fertilization in older plants. *Plant Physiology*, 129(3):1222 – 1231.

Caputi, L., Malnoy, M., Goremykin, V., Nikiforova, S., and Martens, S. (2012). A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. *The Plant Journal*, 69(6):1030 – 1042.

Carels, N. and Bernardi, G. (2000). Two Classes of Genes in Plants. *Genetics*, 154(4):1819 – 1825.

Carvunis, A., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. *Nature*.

Castillo, D. A., Kolesnikova, M. D., and Matsuda, S. P. (2013). An Effective Strategy for Exploring Unknown Metabolic Pathways by Genome Mining. *Journal of the American Chemical Society*, 135(15):5885 – 5894.

CBOL-Plant-Working-Group, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S. C. H., van den Berg, C., Bogarin, D., Burgess, K. S., Cameron, K. M., Carine, M., Chacn, J., Clark, A., Clarkson, J. J., Conrad, F., Devey, D. S., Ford, C. S., Hedderson, T. A., Hollingsworth, M. L., Husband, B. C., Kelly, L. J., Kesanakurti, P. R., Kim, J. S., Kim, Y.-D., Lahaye, R., Lee, H.-L., Long, D. G., Madrin, S., Maurin, O., Meusnier, I., Newmaster, S. G., Park, C.-W., Percy, D. M., Petersen, G., Richardson, J. E., Salazar, G. A., Savolainen, V., Seberg, O., Wilkinson, M. J., Yi, D.-K., and Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31):12794 – 12797.

Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: Molecules, networks, populations*, pages 207 – 232.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of brassica rapa. *PLoS One*, 7(5):e36442.

Christiaens, J. F., Van Mulders, S. E., Duitama, J., Brown, C. A., Ghequire, M. G., De Meester, L., Michiels, J., Wenseleers, T., Voordeckers, K., and Verstrepen, K. J. (2012). Functional divergence of gene duplicates through ectopic recombination. *EMBO reports*, 13(12):1145 – 1151.

Christin, P.-A., Besnard, G., Samaritani, E., Duvall, M. R., Hodkinson, T. R., Savolainen, V., and Salamin, N. (2008). Oligocene {CO2} Decline Promoted {C4} Photosynthesis in Grasses. *Current Biology*, 18(1):37 – 43.

Chu, H. Y., Wegel, E., and Osbourn, A. (2011). From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *The Plant Journal*, 66(1):66 – 79.

Cohen, O., Ashkenazy, H., Burstein, D., and Pupko, T. (2012). Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics*, 28(18):i389 – i394.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938 – 950.

Cook, P. R. (2010). A model for all genomes: the role of transcription factories. *Journal of Molecular Biology*, 395(1):1 – 10.

Cooper, V. S., Vohr, S. H., Wrocklage, S. C., and Hatcher, P. J. (2010). Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Computational Biology*, 6(4):e1000732.

Cope, N. F., Fraser, P., Eskiw, C. H., et al. (2010). The yin and yang of chromatin spatial organization. *Genome Biology*, 11(3):204.

Creevey, C. J. and McInerney, J. O. (2005). Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21(3):390 – 392.

Crombie, W. M. L. and Crombie, L. (1986). Distribution of avenacins A-1, A-2, B-1 and B-2 in oat roots: Their fungicidal activity towards 'Take-all' fungus. *Phytochemistry*, 25(9):2069 – 2073.

Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences*, 105(29):10039 – 10044.

De, S. and Babu, M. M. (2010). Genomic neighbourhood and the regulation of gene expression. *Current Opinion in Cell Biology*, 22(3):326 – 333.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898 – 2903.

Deacon, J. and Mitchell, R. (1985). Toxicity of oat roots, oat root extracts, and saponins to zoospores of *pythium spp.* and other fungi. *Transactions of the British Mycological Society*, 84(3):479 – 487.

DHont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*.

Dick, R., Rattei, T., Haslbeck, M., Schwab, W., Gierl, A., and Frey, M. (2012). Comparative Analysis of Benzoxazinoid Biosynthesis in Monocots and Dicots: Independent Recruitment of Stabilization and Activation Functions. *The Plant Cell*, 24(3):915 – 928.

Drossou, A., Katsiotis, A., Leggett, J., Loukas, M., and Tsakas, S. (2004). Genome and species relationships in genus *avena* based on RAPD and AFLP molecular markers. *Theoretical and Applied Genetics*, 109:48 – 54.

Ducat, D. C., Way, J. C., and Silver, P. A. (2011). Engineering cyanobacteria to generate high-value products. *Trends in biotechnology*, 29(2):95 – 103.

Dutartre, L., Hilliou, F., and Feyereisen, R. (2012). Phylogenomics of the benzoxazinoid biosynthetic pathway of *Poaceae*: gene duplications and origin of the Bx cluster. *BMC Evolutionary Biology*, 12(1):64.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. *Nature cell biology*, 10(9):1106 – 1113.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, 23(1):205 – 211.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792 – 1797.

Eichler, E. E. and Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634):793 – 797.

Ellis, T., Adie, T., and Baldwin, G. (2011). DNA assembly for synthetic biology: from parts to pathways and beyond. *Integrative Biology*, 3(2):109 – 118.

Espinosa-Soto, C. and Wagner, A. (2010). Specialization can drive the evolution of modularity. *PLoS Computational Biology*, 6(3):e1000719.

Facchini, P. J., Bohlmann, J., Covello, P. S., De Luca, V., Mahadevan, R., Page, J. E., Ro, D.-K., Sensen, C. W., Storms, R., and Martin, V. J. (2012). Synthetic biosystems for the production of high-value plant metabolites. *Trends in biotechnology*, 30(3):127 – 131.

Fang, G., Rocha, E. P., and Danchin, A. (2008). Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, 9(1):4.

Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, 338(6108):758 – 767.

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

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397 – 405.

Field, B., Fiston-Lavier, A.-S., Kemen, A., Geisler, K., Quesneville, H., and Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences*, 108(38):16116 – 16121.

Field, B. and Osbourn, A. E. (2008). Metabolic Diversification-Independent Assembly of Operon-Like Gene Clusters in Different Plants. *Science*, 320(5875):543 – 547.

Fischbach, M. A., Walsh, C. T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences*, 105(12):4601 – 4608.

Fondi, M., Emiliani, G., and Fani, R. (2009). Origin and evolution of operons and metabolic pathways. *Research in Microbiology*, 160(7):502 – 512.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., and Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. *Genome Biology*, 10(4):R37.

Freeling, M. (2009). Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology*, 60(1):433 – 453.

Frey, M., Schullehner, K., Dick, R., Fiesselmann, A., and Gierl, A. (2009). Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. *Phytochemistry*, 70(15):1645 – 1651.

Fu, Y.-B. and Williams, D. (2008). AFLP variation in 25 *Avena* species. *Theoretical and Applied Genetics*, 117:333 – 342.

Gachon, C. M., Langlois-Meurinne, M., and Saindrenan, P. (2005). Plant secondary metabolism glycosyltransferases: the emerging functional analysis. *Trends in plant science*, 10(11):542 – 549.

Gao, F. and Zhang, C.-T. (2006). Gc-profile: a web-based tool for visualizing and analyzing the variation of gc content in genomic sequences. *Nucleic acids research*, 34(suppl 2):W686 – W691.

Geisler, K., Hughes, R. K., Sainsbury, F., Lomonossoff, G. P., Rejzek, M., Fairhurst, S., Olsen, C.-E., Motawia, M. S., Melton, R. E., Hemmings, A. M., Bak, S., and Osbourn, A. (2013). Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. *Proceedings of the National Academy of Sciences*, pages E3360 – 3367.

Gelfman, S., Cohen, N., Yearim, A., and Ast, G. (2013). DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon–intron structure. *Genome research*, 23(5):789 – 799.

Gershenzon, J. and Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, 3:408 – 414.

Gibson, D. M. and Krasnoff, S. B. (1999). Exploring the Potential of Biologically Active Compounds from Plants and Fungi. *Biologically Active Natural Products*.

Gilbert, C. and Cordaux, R. (2013). Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biology and Evolution*, 5(5):822 – 832.

Gish, W. (1994). Washington University BLAST (WU BLAST) version 2.0.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178 – D1186.

Goodwin, R. H. and Kavanagh, F. (1948). Fluorescing substances in roots. *Bulletin of the Torrey Botanical Club*, 75(1):1 – 17.

Goodwin, R. H. and Pollock, B. M. (1954). Studies on roots. I. Properties and distribution of fluorescent constituents in avena roots. *Botanical Society of America*, pages 516 – 520.

Gout, J.-F., Kahn, D., Duret, L., et al. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics*, 6(5):e1000944.

Grassi, L. and Tramontano, A. (2011). Horizontal and vertical growth of s. cerevisiae metabolic network. *BMC Evolutionary Biology*, 11(1):301.

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D., and Dunham, M. J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics*, 4(12):e1000303.

Grün, S., Frey, M., and Gierl, A. (2005). Evolution of the indole alkaloid biosynthesis in the genus *Hordeum*: Distribution of gramine and DIBOA and isolation of the benzoxazinoid biosynthesis genes from *Hordeum*. *Phytochemistry*, 66(11):1264 – 1272.

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41:95 – 98.

Hanada, K., Sawada, Y., Kuromori, T., Klausnitzer, R., Saito, K., Toyoda, T., Shinozaki, K., Li, W.-H., and Hirai, M. Y. (2011). Functional compensation of primary and secondary metabolites by duplicate genes in arabidopsis thaliana. *Molecular Biology and Evolution*, 28(1):377 – 382.

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., and Shiu, S.-H. (2008). Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiology*, 148(2):993 – 1003.

Haralampidis, K., Bryan, G., Qi, X., Papadopoulou, K., Bakht, S., Melton, R., and Osbourn, A. (2001). A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots. *Proceedings of the National Academy of Sciences*, 98(23):13431 – 13436.

Hermsen, R., Tans, S., and Ten W., P. R. (2006). Transcriptional regulation by competing transcription factor modules. *PLoS Computational Biology*, 2(12):e164.

Herrin, D. L. and Schmidt, G. W. (1988). Rapid, reversible staining of northern blots prior to hbridization. *Biotechniques*, 6:196 – 200.

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 108(49):19451 – 19452.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754 – 755.

Hurst, L. D., Pál, C., and Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics*, 5(4):299 – 310.

Huson, D. H. and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23(2):254 – 267.

Inagaki, Y.-S., Etherington, G., Geisler, K., Field, B., Dokarry, M., Ikeda, K., Mutsukado, Y., Dicks, J., and Osbourn, A. (2011). Investigation of the potential for triterpene synthesis in rice through genome mining and metabolic engineering. *New Phytologist*, 191(2):432 – 448.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97 – 108.

Irimia, M., Maeso, I., Roy, S. W., and Fraser, H. B. (2013). Ancient *cis*-regulatory constraints and the evolution of genome architecture. *Trends in Genetics*.

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., Bocobza, S. E., Unger, T., Malitsky, S., Finkers, R., Tikunov, Y., Bovy, A., Chikate, Y., Singh, P., Rogachev, I., Beekwilder, J., Giri, A. P., and Aharoni, A. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, 341(6142):175 – 179.

Iyer, V. R. (2012). Nucleosome positioning: bringing order to the eukaryotic genome. *Trends in Cell Biology*, 22(5):250 – 256.

Janga, S. C., Collado-Vides, J., and Babu, M. M. (2008). Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761 – 15766.

Jensen, N. B., Zagrobelny, M., Hjernø, K., Olsen, C. E., Houghton-Larsen, J., Borch, J., Møller, B. L., and Bak, S. (2011). Convergent evolution in biosynthesis of cyanogenic defence compounds in plants and insects. *Nature Communications*, 2:273.

Jiang, W.-k., Liu, Y.-l., Xia, E.-h., and Gao, L.-z. (2013). Prevalent role of gene features in determining evolutionary fates of wgd duplicated genes in flowering plants. *Plant Physiology*, pages 1844 – 1861.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97 – 100.

Joshi, C. P. and Chiang, V. L. (1998). Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases. *Plant Molecular Biology*, 37(4):663 – 674.

Kato-Noguchi, H. (2009). Stress-induced allelopathic activity and momilactone B in rice. *Plant Growth Regulation*, 59(2):153 – 58.

Kato-Noguchi, H. (2011). Barnyard grass-induced rice allelopathy and momilactone B. *Journal of plant physiology*, 168(10):1016 – 1020.

Kato-Noguchi, H., Hasegawa, M., Ino, T., Ota, K., and Kujime, H. (2010). Contribution of momilactone a and b to rice allelopathy. *Journal of Plant Physiology*, 167(10):787 – 791.

Képès, F., Jester, B. C., Lepage, T., Rafiei, N., Rosu, B., and Junier, I. (2012). The layout of a bacterial genome. *FEBS letters*, 586(15):2043 – 2048.

Kersey, P. J., Staines, D. M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J. C., Hughes, D. S. T., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M. D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A., and

Birney, E. (2012). Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Research*, 40(D1):D91 – D97.

Kersting, A. R., Bornberg-Bauer, E., Moore, A. D., and Grath, S. (2012). Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biology and Evolution*, 4(3):316 – 329.

Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H., and Fedorova, N. D. (2010). SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*, 47(9):736 – 741.

Khalil, A. S., Lu, T. K., Bashor, C. J., Ramirez, C. L., Pyenson, N. C., Joung, J. K., and Collins, J. J. (2012). A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, 150(3):647 – 658.

Kliebenstein, D. K. and Osbourn, A. (2012). Making new molecules - evolution of pathways for novel metabolites in plants. *Current Opinion in Plant Biology*, 15(4):415 – 423.

Klironomos, F. D., Berg, J., and Collins, S. (2013). How epigenetic mutations can affect genetic evolution: Model and mechanism. *BioEssays*.

Koonin, E. V. (2009). Evolution of genome architecture. *The International Journal of Biochemistry & Cell Biology*, 41(2):298 – 306.

Koonin, E. V. and Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, 11(7):487 – 498.

Kovács, K., Hurst, L. D., and Papp, B. (2009). Stochasticity in protein levels drives colinearity of gene order in metabolic operons of *escherichia coli*. *PLoS Biology*, 7(5):e1000115.

Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Pea-Rodrguez, L. M., Katsarou, D., Field, B., Osbourn, A. E., and Papadopoulou, K. K. (2013). A metabolic gene cluster in lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytologist*.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1):D1202 – D1210.

Larhlimi, A., Blachon, S., Selbig, J., and Nikoloski, Z. (2011). Robustness of metabolic networks: A review of existing definitions. *Biosystems*, 106(1):1 – 8.

Lawrence, J. G. and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843 – 1860.

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. (2013). Strong Purifying Selection at Synonymous Sites in *D. melanogaster*. *PLoS genetics*, 9(5):e1003527.

Lemons, D. and McGinnis, W. (2006). Genomic evolution of hox gene clusters. *Science*, 313(5795):1918 – 1922.

Lepesheva, G. I. and Waterman, M. R. (2007). Sterol 14-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1770(3):467 – 477.

Lercher, M. J. and Hurst, L. D. (2006). Co-expressed yeast genes cluster over a long range but are not regularly spaced. *Journal of Molecular Biology*, 359(3):825 – 831.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178 – 2189.

Li, W.-T., Peng, Y.-Y., Wei, Y.-M., Baum, B. R., and Zheng, Y.-L. (2009). Relationships among *Avena* species as revealed by consensus chloroplast simple sequence repeat (ccSSR) markers. *Genetic Resources and Crop Evolution*, 56(4):465 – 480.

Lim, H. N., Lee, Y., and Hussein, R. (2011). Fundamental relationship between operon organization and gene expression. *Proceedings of the National Academy of Sciences*, 108(26):10626 – 10631.

Liscombe, D. K., Louie, G. V., and Noel, J. P. (2012). Architectures, mechanisms and molecular evolution of natural product methyltransferases. *Natural Product Reports*, 29(10):1238 – 1250.

Lobkovsky, A. E. and Koonin, E. V. (2012). Replaying the tape of life: quantification of the predictability of evolution. *Frontiers in Genetics*, 3.

Loskutov, I. G. (2008). On evolutionary pathways of *avena* species. *Genetic Resources and Crop Evolution*, 55:211 – 220.

Lukens, L. N., Pires, J. C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of sequence loss and cytosine methylation within a population of newly resynthesized brassica napus allopolyploids. *Plant Physiology*, 140(1):336 – 348.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459 – 473.

Madan Babu, M., Janga, S. C., de Santiago, I., and Pombo, A. (2008). Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current Opinion in Genetics & Development*, 18(6):571 – 582.

Maizel, J. V., Burkhardt, H. J., and Mitchell, H. K. (1964). Avenacin, an Antimicrobial Substance Isolated from Avena sativa. I. Isolation and Antimicrobial Activity*. *Biochemistry*, 3(3):424 – 426.

Martin, F. J. and McInerney, J. O. (2009). Recurring cluster and operon assembly for phenylacetate degradation genes. *BMC Evolutionary Biology*, 9(1):36.

Masel, J. and Trotter, M. V. (2010). Robustness and evolvability. *Trends in genetics: TIG*, 26(9):406.

Matsuba, Y., Nguyen, T. T., Wiegert, K., Falara, V., Gonzales-Vigil, E., Leong, B., Schfer, P., Kudrna, D., Wing, R. A., Bolger, A. M., Usadel, B., Tissier, A., Fernie, A. R., Barry, C. S., and Pichersky, E. (2013). Evolution of a complex locus for terpene biosynthesis in solanum. *The Plant Cell*.

Matsuno, M., Compagnon, V., Schoch, G. A., Schmitt, M., Debayle, D.and Bassard, J.-E., Pollet, B., Hehn, A., Heintz, D., Ullmann, P., Lapierre, C., Bernier, F.and Ehlting, J., and Werck-Reichhart, D. (2009). Evolution of a Novel Phenolic Pathway for Pollen Development. *Science*, 325(5948):1688 – 1692.

McGary, K. L., Slot, J. C., and Rokas, A. (2013). Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. 110(28):11481 – 11486.

Medema, M. H., Takano, E., and Breitling, R. (2013). Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Molecular Biology and Evolution*, 30(5):1218 – 1223.

Mert-Turk, F., Egesel, C., and Gul, M. (2005). Avenacin A-1 content of some local oat genotypes and the in vitro effect of avenacins on several soil-borne fungal pathogens of cereals. *Turkish Journal of Agriculture and Forestry*, 29:157 – 164.

Milo, R. and Last, R. L. (2012). Achieving diversity in the face of constraints: lessons from metabolism. *Science*, 336(6089):1663 – 1667.

Mochida, K., Yoshida, T., Sakurai, T., and Ogihara, Y.and Shinozaki, K. (2009). TriFLDB: A Database of Clustered Full-Length Coding Sequences from Triticeae with Applications to Comparative Grass Genomics. *Plant Physiology*, 150(3):1135 – 1146.

Morikawa, T. and Nishihara, M. (2009). Genomic and polyploid evolution in genus *Avena* as revealed by RFLPs of repeated DNA sequences. *Genes & Genetic Systems*, 84(3):199 – 208.

Morrone, D., Hillwig, M., Mead, M., Lowry, L., Fulton, D., and Peters, R. (2011). Evident and latent plasticity across the rice diterpene synthase family with potential implications for the evolution of diterpenoid metabolism in the cereals. *Biochemical Journal*, 435:589 – 595.

Mugford, S. and Osbourn, A. (2010). Evolution of serine carboxypeptidase-like acyltransferases in the monocots. *Plant Signaling & Behavior*, 5:193 – 195.

Mugford, S. T., Louveau, T., Melton, R., Qi, X., Bakht, S., Hill, L., Tsurushima, T., Honkanen, S., Rosser, S. J., Lomonossoff, G. P., et al. (2013). Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. *The Plant Cell*, 25(3):1078 – 1092.

Mugford, S. T. and Milkowski, C. (2012). Chapter Fourteen - Serine Carboxypeptidase-Like Acyltransferases from Plants. In *Natural Product Biosynthesis by Microorganisms and Plants, Part B*, volume 516 of *Methods in Enzymology*, pages 279 – 297. Academic Press.

Mugford, S. T., Qi, X., Bakht, S., Hill, L., Wegel, E., Hughes, R. K., Papadopoulou, K., Melton, R., Philo, M., Sainsbury, F., Lomonossoff, G. P., Roy, A. D., Goss, R. J., and Osbourn, A. (2009). A Serine Carboxypeptidase-Like Acyltransferase Is Required for Synthesis of Antimicrobial Compounds and Disease Resistance in Oats. *The Plant Cell*, 21(8):2473 – 2484.

Murat, F., Van de Peer, Y., and Salse, J. (2012). Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes. *Genome biology and evolution*, 4(9):917 – 928.

Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., Messing, J., and Salse, J. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research*, 20(11):1545 – 1557.

Muro, E. M., Mah, N., Moreno-Hagelsieb, G., and Andrade-Navarro, M. A. (2011). The pseudogenes of mycobacterium leprae reveal the functional relevance of gene order within operons. *Nucleic Acids Research*, 39(5):1732 – 738.

Mylona, P., Owatworakit, A., Papadopoulou, K., Jenner, H., Qin, B., Findlay, K., Hill, L., Qi, X., Bakht, S., Melton, R., and Osbourn, A. (2008). Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat. *The Plant Cell*, 20(1):201 – 212.

Nakagawa, A., Minami, H., Kim, J.-S., Koyanagi, T., Katayama, T., Sato, F., and Kumagai, H. (2011). A bacterial platform for fermentative production of plant alkaloids. *Nature communications*, 2:326.

Nannapaneni, K., Ben-Shahar, Y., Keen, H. L., Welsh, M. J., Casavant, T. L., and Scheetz, T. E. (2013). Computational identification of operon-like transcriptional loci in eukaryotes. *Computers in Biology and Medicine*.

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338(6105):384 – 387.

Nemoto, T., Okada, A., Okada, K., Shibuya, N., Toyomasu, T., Nojiri, H., and Yamane, H. (2007). Promoter analysis of the rice stemar-13-ene synthase gene *OsDTC2*, which is involved in the biosynthesis of the phytoalexin oryzalexin S. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1769(11):678 – 683.

Nikoloudakis, N. and Katsiotis, A. (2008). The origin of the C-genome and cytoplasm of *Avena* polyploids. *Theoretical and Applied Genetics*, 117(2):273 – 281.

Nikoloudakis, N., Skaracis, G., and Katsiotis, A. (2008). Evolutionary insights inferred by molecular analysis of the ITS1-5.8S-ITS2 and IGS *Avena sp.* sequences. *Molecular Phylogenetics and Evolution*, 46(1):102 – 115.

Nomura, T., Ishihara, A., Imaishi, H., Ohkawa, H., Endo, T. R., and Iwamura, H. (2003). Rearrangement of the genes for the biosynthesis of benzoxazinones in the evolution of triticeae species. *Planta*, 217(5):776 – 782.

Norris, V. and Merieau, A. (2013). Plasmids as scribbling pads for operon formation and propagation. *Research in Microbiology*, pages 779 – 787.

Ober, D. (2010). Gene duplications and the time thereafter–examples from plant secondary metabolism. *Plant Biology*, 12(4):570 – 577.

Ohno, S. (1970). Evolution by gene duplication. *London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.*

Okada, A.and Okada, K., Miyamoto, K., Koga, J., Shibuya, N., Nojiri, H., and Yamane, H. (2009). OsTGAP1, a bZIP transcription factor, coordinately regulates the inductive production of diterpenoid phytoalexins in rice. *Journal of Biological Chemistry*, 284(39):26510 – 26518.

Osbourn, A. (2010a). Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant physiology*, 154(2):531 – 535.

Osbourn, A. (2010b). Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. *Plant Physiology*, 154(2):531 – 535.

Osbourn, A., Clarke, B., Lunness, P., Scott, P., and Daniels, M. (1994). An oat species lacking avenacin is susceptible to infection by gaeumannomyces graminis var. tritici. *Physiological and Molecular Plant Pathology*, 45(6):457 – 467.

Osbourn, A. E. and Field, B. (2009). Operons. *Cellular and Molecular Life Sciences*, 66(23):3755 – 3775.

O'Sullivan, J., Sontam, D., Grierson, R., and Jones, B. (2009). Repeated elements coordinate the spatial organization of the yeast genome. *Yeast*, 26(2):125 – 138.

Owatworakit, A., Townsend, B., Louveau, T., Jenner, H., Rejzek, M., Hughes, R. K., Saalbach, G., Qi, X., Bakht, S., Roy, A. D., Mugford, S. T., Goss, R. J. M., Field, R. A., and Osbourn, A. (2013). Glycosyltransferases from Oat (Avena) Implicated in the Acylation of Avenacins. *Journal of Biological Chemistry*, 288(6):3696 – 3704.

Pál, C. and Hurst, L. D. (2004). Evidence against the selfish operon theory. *Trends in Genetics*, 20(6):232 – 234.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., and Osbourn, A. E. (1999). Compromised disease resistance in saponin-deficient plants. *Proceedings of the National Academy of Sciences*, 96(22):12923 – 12928.

Park, S. G. and Choi, S. S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evolutionary Biology*, 10(1):241.

Peng, Y.-Y., Baum, B. R., Ren, C.-Z., Jiang, Q.-T., Chen, G.-Y., Zheng, Y.-L., and Wei, Y.-M. (2010a). The evolution pattern of rDNA ITS in *Avena* and phylogenetic relationship of the *Avena* species (*Poaceae: Aveneae*). *Hereditas*, 147(5):183 – 204.

Peng, Y.-Y., Wei, Y.-M., Baum, B. R., Jiang, Q.-T., Lan, X.-J., Dai, S.-F., and Zheng, Y.-L. (2010b). Phylogenetic investigation of *Avena* diploid species and the maternal genome donor of *Avena* polyploids. *Taxon*, 59(5):1472 – 1482.

Peng, Y.-Y., Wei, Y.-M., Baum, B. R., Yan, Z.-H., Lan, X.-J., Dai, S.-F., and Zheng, Y.-L. (2010c). Phylogenetic inferences in *Avena* based on analysis of FL intron2 sequences. *TAG Theoretical and Applied Genetics*, 121(5):985 – 1000.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera-A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605 – 1612.

Phillips, D. R., Rasbery, J. M., Bartel, B., and Matsuda, S. (2006). Biosynthetic diversity in plant triterpene cyclization. *Current Opinion in Plant Biology*, 9(3):305 – 314.

Pichersky, E. and Lewinsohn, E. (2011). Convergent Evolution in Plant Specialized Metabolism. *Annual Review of Plant Biology*, 62(1):549 – 566.

Posada, D. and Crandall, K. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4):580 – 601.

Price, M. N., Arkin, A. P., and Alm, E. J. (2006). The life-cycle of operons. *PLoS Genetics*, 2(6):e96.

Price, M. N., Huang, K. H., Arkin, A. P., and Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Research*, 15(6):809 – 819.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., and Osbourn, A. (2004). A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):8233 – 8238.

Qi, X., Bakht, S., Qin, B., Leggett, M., Hemmings, A., Mellon, F., Eagles, J., Werck-Reichhart, D., Schaller, H., Lesot, A., Melton, R., and Osbourn, A. (2006). A different function for a member of an ancient and highly conserved cytochrome P450 family: From essential sterols to plant defense. *Proceedings of the National Academy of Sciences*, 103(49):18848 – 18853.

Quintanar, A., Castroviejo, S., and Cataln, P. (2007). Phylogeny of the tribe *Aveneae (Pooideae, Poaceae)* inferred from plastid trnT-F and nuclear ITS sequences. *American Journal of Botany*, 94(9):1554 – 1569.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Racolta, S., Juhl, P. B., Sirim, D., and Pleiss, J. (2012). The triterpene cyclase protein family: A systematic analysis. *Proteins: Structure, Function, and Bioinformatics*, 80(8):2009 – 2019.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551 – 1555.

Ray, J. C. J. and Igoshin, O. A. (2012). Interplay of gene expression noise and ultrasensitive dynamics affects bacterial operon organization. *PLoS computational Biology*, 8(8):e1002672.

Renner, S. S. and Bellot, S. (2012). Horizontal gene transfer in eukaryotes: Fungi-to-plant and plant-to-plant transfers of organellar DNA. In *Genomics of Chloroplasts and Mitochondria*, pages 223 – 235. Springer.

Rodgers-Melnick, E., Mane, S. P., Dharmawardhana, P., Slavov, G. T., Crasta, O. R., Strauss, S. H., Brunner, A. M., and DiFazio, S. P. (2012). Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research*, 22(1):95 – 105.

Rodionov, A., Tyupa, N., Kim, E., Machs, E., and Loskutov, I. (2005). Genomic configuration of the autotetraploid oat species *Avena macrostachya* inferred from comparative analysis of ITS1 and ITS2 sequences: on the oat karyotype evolution during the early events of the *Avena* species divergence. *Russian Journal of Genetics*, 41(5):518 – 528.

Rodrigues, J. F. M. and Wagner, A. (2009). Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Computational Biology*, 5(12):e1000613.

Roy, S. and Irimia, M. (2009). Mystery of intron gain: new data and new models. *Trends in Genetics*, 25(2):67 – 73.

Saarela, J. M., Liu, Q., Peterson, P. M., Soreng, R. J., and Paszko, B. (2010). Phylogenetic of the grass 'Aveneae-Type Plastid DNA Clade' (*poaceae: pooideae, poeae*) based on plastid and nuclear ribosomal DNA sequence data. *Diverswity, Phylogeny, and Evolution in the monocotyledons*.

Sainsbury, F., Thuenemann, E. C., and Lomonossoff, G. P. (2009). pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant biotechnology journal*, 7(7):682 – 693.

Saito, K. (2013). Phytochemical genomics - a new trend. *Current opinion in plant biology*, pages 373 – 380.

Sawai, S. and Saito, K. (2011). Triterpenoid biosynthesis and engineering in plants. *Frontiers in Plant Science*, 2.

Schnable, J. C., Wang, X., Pires, J. C., and Freeling, M. (2012). Escape from preferential retention following repeated whole genome duplications in plants. *Frontiers in Plant Science*, 3.

Schullehner, K., Dick, R., Vitzthum, F., Schwab, W., Brandt, W., Frey, M., and Gierl, A. (2008). Benzoxazinoid biosynthesis in dicot plants. *Phytochemistry*, 69(15):2668 – 2677.

Serres-Giardi, L., Belkhir, K., David, J., and Glmin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *The Plant Cell*, 24(4):1379 – 1397.

Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic acids research*, 41(4):2073 – 2094.

Shelukhina, O. Y., Badaeva, E., Brezhneva, T., Loskutov, I., and Pukhalsky, V. (2008a). Comparative analysis of diploid species of *Avena L.* Using cytogenetic and biochemical markers: *Avena pilosa* MB and *A. clauda Dur. Russian Journal of Genetics*, 44(9):1087 – 1091.

Shelukhina, O. Y., Badaeva, E., Brezhneva, T., Loskutov, I., and Pukhalsky, V. (2008b). Comparative analysis of diploid species of *Avena L.* Using cytogenetic and biochemical markers: *Avena pilosa* MB and *A. clauda Dur. Russian Journal of Genetics*, 44(9):1087 – 1091.

Siddiqui, M. S., Thodey, K., Trenchard, I., and Smolke, C. D. (2012). Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools. *FEMS yeast research*, 12(2):144 – 170.

Sikosek, T., Chan, H. S., and Bornberg-Bauer, E. (2012). Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proceedings of the National Academy of Sciences*, 109(37):14888 – 14893.

Smolke, C. D. (2009). Building outside of the box: iGEM and the BioBricks Foundation. *Nature biotechnology*, 27(12):1099 – 1102.

Soderlund, C., Bomhoff, M., and Nelson, W. M. (2011). SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic acids research*, 39(10):e68 – e68.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688 – 2690.

Stehle, F., Brandt, W., Milkowski, C., and Strack, D. (2006). Structure determinants and substrate recognition of serine carboxypeptidase-like acyltransferases from plant secondary metabolism. *FEBS Letters*, 580(27):6366 – 6374.

Sue, M., Nakamura, C., and Nomura, T. (2011). Dispersed Benzoxazinone Gene Cluster: Molecular Characterization and Chromosomal Localization of Glucosyltransferase and Glucosidase Genes in Wheat and Rye. *Plant Physiology*, 157(3):985 – 997.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl 2):W609 – W612.

Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B., and Peters, R. J. (2009). CYP76M7 is an ent-cassadiene C11α-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *The Plant Cell*, 21(10):3315 – 3325.

Takos, A. M., Knudsen, C., Lai, D., Kannangara, R., Mikkelsen, L., Motawia, M. S., Olsen, C. E., Sato, S., Tabata, S., Jorgensen, K., Moller, B. L., and Rook, F. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *The Plant Journal*, 68:273 – 286.

Takuno, S. and Gaut, B. S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, 110(5):1797 – 1802.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28(10):2731 – 2739.

Tang, H., Bowers, J. E., Wang, X., and Paterson, A. H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences*, 107(1):472 – 477.

Tatarinova, T., Elhaik, E., and Pellegrini, M. (2013). Cross-Species Analysis of Genic GC3 Content and DNA Methylation Patterns. *Genome Biology and Evolution*, 5(8):1443 – 1456.

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., and Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC genomics*, 11(1):308.

Thoma, R., Schulz-Gasch, T., D'Arcy, B., Benz, J., Aebi, J., Dehmlow, H., Hennig, M., Stihle, M., and Ruf, A. (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature*, 432(7013):118 – 122.

Thomas, S., Bonello, P., Lipps, P. E., and Boehm, M. J. (2006). Avenacin Production in Creeping Bentgrass *(Agrostis stolonifera)* and Its Influence on the Host Range of Gaeumannomyces graminis. *Plant Disease*, 90(1):33 – 38.

Treangen, T. J. and Rocha, E. P. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics*, 7(1):e1001284.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725 – 732.

Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of Genome Duplications from Age Distributions Revisited. *Molecular Biology and Evolution*, 30(1):177 – 190.

Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., and Verstrepen, K. J. (2012). Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biology*, 10(12):e1001446.

Wagner, A. (2008). Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91 – 100.

Wang, C.-T., Ho, C.-H., Hseu, M.-J., and Chen, C.-M. (2010). The subtelomeric region of the arabidopsis thaliana chromosome iiir contains potential genes and duplicated fragments from other chromosomes. *Plant Molecular Biology*, 74(1-2):155 – 166.

Wang, Q., Hillwig, M. L., and Peters, R. J. (2011). CYP99A3: functional identification of a diterpene oxidase from the momilactone biosynthetic gene cluster in rice. *The Plant Journal*, 65(1):87 – 95.

Wang, Y. (2013). Locally-duplicated ohnologs evolve faster than non-locally-duplicated ohnologs in Arabidopsis and rice. *Genome Biology and Evolution*.

Wang, Z. and Zhang, J. (2009). Abundant indispensable redundancies in cellular metabolic networks. *Genome Biology and Evolution*, 1:23.

Wegel, E., Koumproglou, R., Shaw, P., and Osbourn, A. (December 2009). Cell Type-Specific Chromatin Decondensation of a Metabolic Gene Cluster in Oats. *The Plant Cell*, 21(12):3926 – 3936.

Wendt, K. U., Poralla, K., and Schulz, G. E. (1997). Structure and function of a squalene cyclase. *Science*, 277(5333):1811 – 1815.

Weng, J.-K., Philippe, R. N., and Noel, J. P. (2012). The rise of chemodiversity in plants. *Science*, 336(6089):1667 – 1670.

Wicker, T., Buchmann, J. P., and Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Research*, 20(9):1229 – 1237.

Wijchers, P. and de Laat, W. (2011). Genome organization influences partner selection for chromosomal rearrangements. *Trends in Genetics*, 27(2):63 – 71.

Winterfeld, G., Doring, E., and Roser, M. (2009). Chromosome evolution in wild oat grasses *(aveneae)* revealed by molecular phylogeny. *Genome*, 52(4):361 – 380.

Winzer, T., Gazda, V., He, Z., Kaminski, F., Kern, M., Larson, T. R., Li, Y., Meade, F., Teodor, R., Vaistij, F. E., et al. (2012). A papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, 336(6089):1704 – 1708.

Wu, D.-D. and Zhang, Y.-P. (2013). Evolution and function of de novo originated genes. *Molecular Phylogenetics and Evolution*, pages 541 – 545.

Wu, X. and Qi, X. (2010). Genes encoding hub and bottleneck enzymes of the arabidopsis metabolic network preferentially retain homeologs through whole genome duplication. *BMC Evolutionary biology*, 10(1):145.

Wu, Z.-Q. and Ge, S. (2012). The phylogeny of the {BEP} clade in grasses revisited: Evidence from the whole-genome sequences of chloroplasts. *Molecular Phylogenetics and Evolution*, 62(1):573 – 578.

Xu, P. and Koffas, M. A. (2013). Assembly of Multi-gene Pathways and Combinatorial Pathway Libraries Through ePathBrick Vectors. *Synthetic Biology*, 1073:107 – 129.

Xue, C., Huang, R., Liu, S.-Q., and Fu, Y.-X. (2010). Recombination facilitates neofunctionalization of duplicate genes via originalization. *BMC Genetics*, 11(1):46.

Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., OMaille, P., Osbourn, A., and Qi, X. (2012). Divergent evolution of oxidosqualene cyclases in plants. *New Phytologist*, 193(4):1022 – 1038.

Yang, Z. (2006). *Computational molecular evolution*, volume 284. Oxford University Press Oxford.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586 – 1591.

Yang, Z. and dos Reis, M. (2011). Statistical Properties of the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*, 28(3):1217 – 1228.

Yeaman, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences*, 110(19):E1743 – E1751.

Yonekura-Sakakibara, K. and Hanada, K. (2011). An evolutionary view of functional diversity in family 1 glycosyltransferases. *The Plant Journal*, 66(1):182 – 193.

Zhang, Y. (2008). I-TASSER server for protein 3d structure prediction. *BMC Bioinformatics*, 9(1):40.

Zhang, Y., Wu, Y., Liu, Y., and Han, B. (2005). Computational Identification of 69 Retroposons in Arabidopsis. *Plant Physiology*, 138(2):935 – 948.

Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H., and Long, M. (2011). A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. *Bioinformatics*, 27(13):1749 – 1753.

Zhu, Y., Du, P., and Nakhleh, L. (2012). Gene duplicability-connectivity-complexity across organisms and a neutral evolutionary explanation. *PloS One*, 7(9):e44491.