Plant genome mining for triterpene biosynthetic genes and gene clusters

A thesis submitted to the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Charlotte Heather Emily Owen

February 2020

Abstract

Plant specialised metabolites are highly diverse in their functions and chemistries. The discovery of plant biosynthetic gene clusters (BGCs) and the rapidly increasing volume of sequence data available for analysis provides a timely opportunity for wide, comprehensive analyses of BGCs across plants. Triterpenes were chosen as exemplars for this, given the solid foundation of established literature and the existence of powerful characterisation platforms to permit an iterative synthetic biology approach. After an assessment of current BGC mining tools, key limitations were identified regarding accuracy and specificity of putative enzyme and pathway classifiers, as well as in variation of genome quality. Many of these limitations were overcome through the creation of systematic tools for locating, classifying and predicting the function of three key triterpene enzyme families: oxidosqualene cyclases (OSCs), cytochrome P450s and glycosyl-transferases. The generation of these tools represent a step-change in our ability to effectively analyse large volumes of sequence data. In the application of these tools, a wide range of data were generated to explore the evolutionary patterns of these families in the Viridiplantae, across a taxonomic range an order of magnitude greater than previous studies. The dynamic and diverse nature of triterpene biosynthetic enzyme evolution was observed, and the methodologies validated by comparison to known biosynthetic pathways and gene clusters. These data, when combined with comprehensive enrichment analysis of gene families co-located with OSCs, have provided a wealth of options for future study. These include: assessing if variation in repertoires of key enzyme subfamilies between plant clades impacts their biosynthetic potential, designer metabolite synthesis via the use of rigorous synthetic biology approaches, assessing non-biosynthetic genes as potential components of BGCs and exploring the space between entirely clustered and non-clustered biosynthetic pathways to build a cohesive model for plant gene organisation in the context of specialised metabolism.

Acknowledgements

Firstly, my sincere gratitude to Anne Osbourn for the opportunity and freedom she has given me throughout this PhD, her astute guidance and insight, as well as her kindness, patience and generosity over some difficult years. The whole Osbourn group, past and present, has my sincerest thanks for their thoughtfulness, openness and joy in approaching the many challenges of scientific research. In particular I wish to thank Ancheng Huang, Thomas Louveau, Zhenhua Liu, Jenny Jo, Anastasia Orme, James Reed and Michael Stephenson for their vast and varied expertise regarding the many facets of this PhD, their good friendship and their kind support. I give special thanks to Hannah Hodgson, who has been utterly steadfast in her encouragement and acceptance, and who has given me so much strength to keep going.

I thank the Rotation PhD Programme and the John Innes Foundation for providing me with the freedom to access so much incredible science across the Norwich Research Park, especially Diane Saunders and Barrie Wilkinson for their supervision and advice during the rotation projects, and Andy Truman for his generous advice and time given in examining this thesis. I give my gratitude to Stephen Bornemann in particular, whose composed, kind and sincere counsel I have appreciated immensely. Across the NRP, I also thank Sophie Harrington, Catherine Jacott and Jack Rhodes for their friendship and much needed coffee breaks across the years.

I would like to acknowledge Marnix Medema and Neil Hall, who have both provided valuable insight at various stages during this PhD. My thanks too to those teachers and academics who have inspired and encouraged me over the years, including Steve Cook, Steven Longstaff and Mary Maunder.

The wider support staff at the JIC must also be thanked, who have helped me both indirectly and directly to complete this thesis, including metabolite services, CiS, the GSO, HR and the IT, stores and finances departments. In particular, I thank Lionel Hill for his continual kindness and Sam Hicks for their solidarity and reassurance when I needed it most.

I shall forever be grateful for the NHS, without which neither this thesis nor the author would likely have made it this far. In particular, I thank all the staff at Bury North IDT, Access and Assessment Suffolk and Wedgewood House. My heartfelt thanks especially go to both Simon Arthur and Vanessa Crawford, two clinicians who dealt with me at critical junctures and who genuinely surprised me with their willingness and drive to help. My gratitude also to Jude Adcock, who has listened, and listened, and listened.

My thanks and love to my family and friends, new and old. To my parents, Julie and Andy, for their love and care. To Mel and Dave, for their continued acceptance and warmth. To Wrenna Robson, Rebecca Morgan and Tam Blaxter, whose solidarity and kindness has been absolute and who have taught me so much. To Joseph Greenwood and Maxwell Phillips, whose humour and company is always uplifting. To Stephanie; I will never truly be able say how much their support, care and love over the last decade and more has helped me live, survive and grow. I shall remain so incredibly grateful for their friendship for all the decades to come.

Finally, my love to all those we have lost, and to all those that yet remain to fight on. For acceptance, for kindness and for peace. For hope, warmth and safety in the darkness. For the joy of life, and the bounty it contains.

Contents

List of Figures

List of Tables

Abbreviations

2OG	2-oxoglutarate
AAA	ATPases associated with diverse cellular activities
ACS	ancestral cycloartenol synthase
ALSL	ancestral lanosterol synthase-like
AT	acyltransferase
BAS	beta-amyrin synthase
BGC	biosynthetic gene cluster
CAS	cycloartenol synthase
CAZy	carbohydrate active enzyme
CCS	cucurbitadienol synthase
CR	catalytic residue
CYP	cytochrome P450
EC	Enzyme Commission
FDR	false discovery rate
GO	gene ontology
GT	glycosyl-transferase
GT1	family 1 glycosyl-transferase
MITE	miniature inverted transposable element
MT	methyltransferase
MVA	mevalonate
OSC	oxidosqualene cyclase
PCC	Pearson correlation coefficient
pHMM	profile Hidden Markov Model
PPR	pentatricopeptide repeats
PSPG	plant secondary product glycosyltransferase
PT	prenyltransferase
SCPL	serine carboxypeptidase-like
SDR	specificity determining residue
SOM	self-organising map
SSR	sugar donor specific residue
TE	transposable element
TPS	terpene synthase
UGT	UDP-dependent glycosyl-transferase

Chapter 1. General introduction

1.1 Finding and using plant specialised metabolites

The chemistries and functions of plant natural products are incredibly diverse and complex. Specialised metabolites form a foundational part of plants' ability to interact with the biota around them, such as in the protection against pathogens, discouraging feeding, pigmentation and inter- and intra-species signalling [1–6]. The myriad uses plants have to humanity are often due to such specialised metabolites and have been utilised in a huge variety of ways throughout history.

A large component of this benefit is from the medicinal activities of specialised metabolites. Plants have been sought for use as medicines by pre-historic humanity. Indeed, animals other than humans are observed to 'self-medicate' by the consumption or application of plant material, from chimpanzees chewing on the leaves of *Vernonia amygdalina* during rainy seasons to reduce infections [7], to 'woolly bear' caterpillars (*Grammia incorrupta*) which selectively consume leaves high in alkaloids when endoparasitised by flies [8,9]. For humans, one of the earliest medicinal texts is the Ebers Papyrus (c. 1500 BCE), which identifies numerous plants with particular utility, such as poppies and nightshade for use as an anaesthetic, liquorice as an expectorant, various plants to repel insects and *Aloe* species to treat burns and skin irritation [10]. Beyond medicine, plant specialised metabolites have historically been used as dyes (such as red madder, blue woad and yellow weld) [11] and soaps [12]. *Quillaja saponaria* and *Saponaria officinalis* were used as traditional detergents for washing fabrics in South America and Europe, respectively [12–15].

With the birth of agriculture, domestication and plant breeding, the production or inhibition of plant specialised metabolites has been selected for. For example, domestic species of Cucurbitaceae, such as melons and cucumbers, have been bred to move the production of bitter-tasting compounds from the fruits to the leaves [16]. A broad trend in crop domestication is the reduction in plant specialised metabolites used for defence against pathogens and insects in favour of harvestability and yield [17,18]. As scientific progress has allowed the study of genetics, genomics and refined metabolite analysis, increasingly detailed approaches to understanding and manipulating specialised metabolic pathways have been developed. This includes the production of foods with increased concentrations of beneficial compounds [19,20] and the heterologous production of plant derived medicinal compounds for large scale production [21,22].

1.2 Modern metabolic and synthetic biology

With the ever-reducing cost of DNA sequencing technology and the computational capacity for assembly of highly complex genomes, the volume of genetic sequence data available for analysis is unprecedented. Figure 1.1 shows the cumulative growth in submissions of whole plant genomes to the NCBI genome database (www.ncbi.nlm.nih.gov/genome/) over the last two decades, and the abundance of transcriptomic data is orders of magnitude greater than that of genomes. In addition to individual labs being able to sequence plant species of interest [23], large scale sequencing projects of multiple species are also underway, such as the 10,000 plant genomes project which is planned to be completed by 2023 [24].

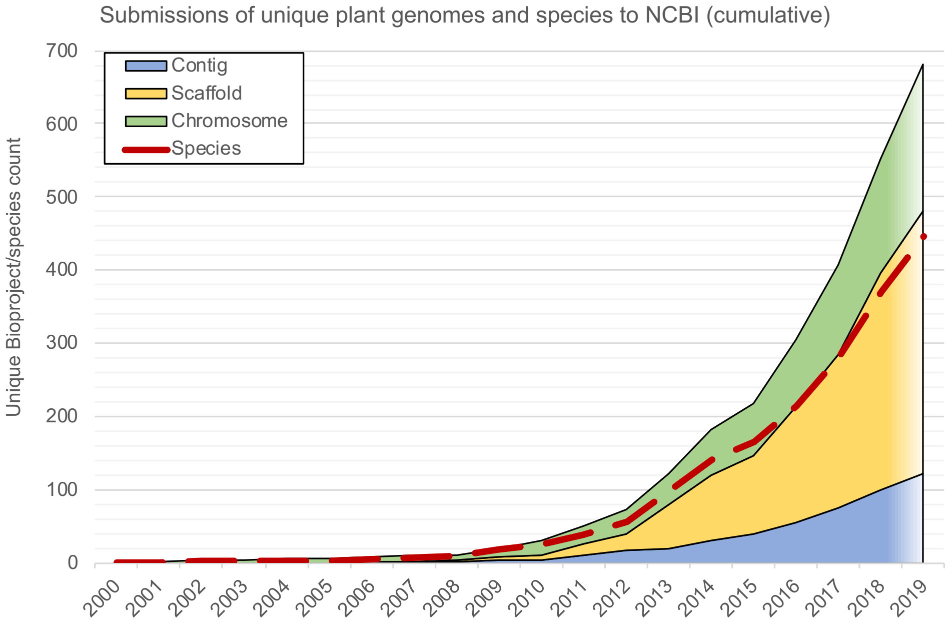


Figure 1.1 **Cumulative growth of plant genome sequences in the NCBI genome database** Reduction of cost in genome sequence technology has allowed individual research groups to fully sequence a given plant genome and consortia to be able to sequences hundreds to thousands of genomes. The volume of data now being generated is such that high-throughput tools are required to handle them effectively.

In the context of plant specialised metabolism, where interest is generally in a set of key biosynthetic gene families and the ancillary genes involved their regulation, these data present a number of key opportunities. The first is in understanding the evolution of such gene families of interest, by leveraging the broad range of species with available sequence data and comparing how genes have diversified and changed across evolutionary time. Projects that have set out to sequence species of plant taxa generally underrepresented in public resources, such as the 1,000 plants (1KP) transcriptomes project [25], have especially increased the power of broad-scale evolutionary analyses.

Secondly, the wealth of sequence data provides material for the generation of tools to classify target gene families, to predict their biosynthetic activity and to select candidates with potentially useful activities for further study. Biosynthetic enzymes often have complex relationships between their sequence, structure and function to achieve the completion of nuanced chemical reactions, so large datasets are often useful to parse out the relevant information. Furthermore, cross-reference with natural products databases and integration with high-throughput analytical platforms for the rapid characterisation of candidate enzymes increases the power of such predictive tools dramatically, in allowing the feedback of validation and testing [26].

To summarise, the scale of sequence data currently available is so large as to require high-throughput, systematic tools and analyses to effectively utilise it. One aim is to predict the activities of target gene family, characterise enzyme activity and subsequently verify and validate the predictive tools used. From a metabolic engineering perspective, an ultimate aim is to be able to make target molecules 'on-demand'. Given this, a synthetic biology methodology in metabolic biology is evidently suitable, where systematic approaches are made towards defining and overcoming challenges in a 'design-build-test' cycle [26]. Of course, progress in understanding the evolutionary processes of natural product genetics also assists the engineering goals, and *vice versa*.

1.3 Plant biosynthetic gene clusters

Given the huge diversity of plant specialised metabolites, their biosynthetic pathways can be highly complex, requiring the involvement of numerous specific, fine-tuned reactions [27,28]. Furthermore, the enzyme families that catalyse such reactions are often members of very large families, the genes of which can be found in their hundreds in a given plant genome [29,30]. The *in planta* roles of these metabolites often require tight spatio-temporal regulation, such as in response to specific elicitors or production in specific tissues [31]. These factors can combine to hinder our ability to rapidly find biosynthetic candidates.

However, the detection of biosynthetic gene clusters (BGCs) in plants has opened a new route for gene discovery. In plant BGCs, coregulated genes for a specific pathway are found co-located in the genome [32–34]. Such a phenomenon therefore gives researchers another dimension to consider when mining sequence data for target genes, which can be combined with co-expression data and sequence-based predictive tools. In this way, much can be borrowed from the advances made in microbial BGC mining and characterisation, and various tools have been developed in recent years for mining plant genomes for BGCs. Examples of characterised plant BGCs and the compounds they produce are shown in Figure 1.2.

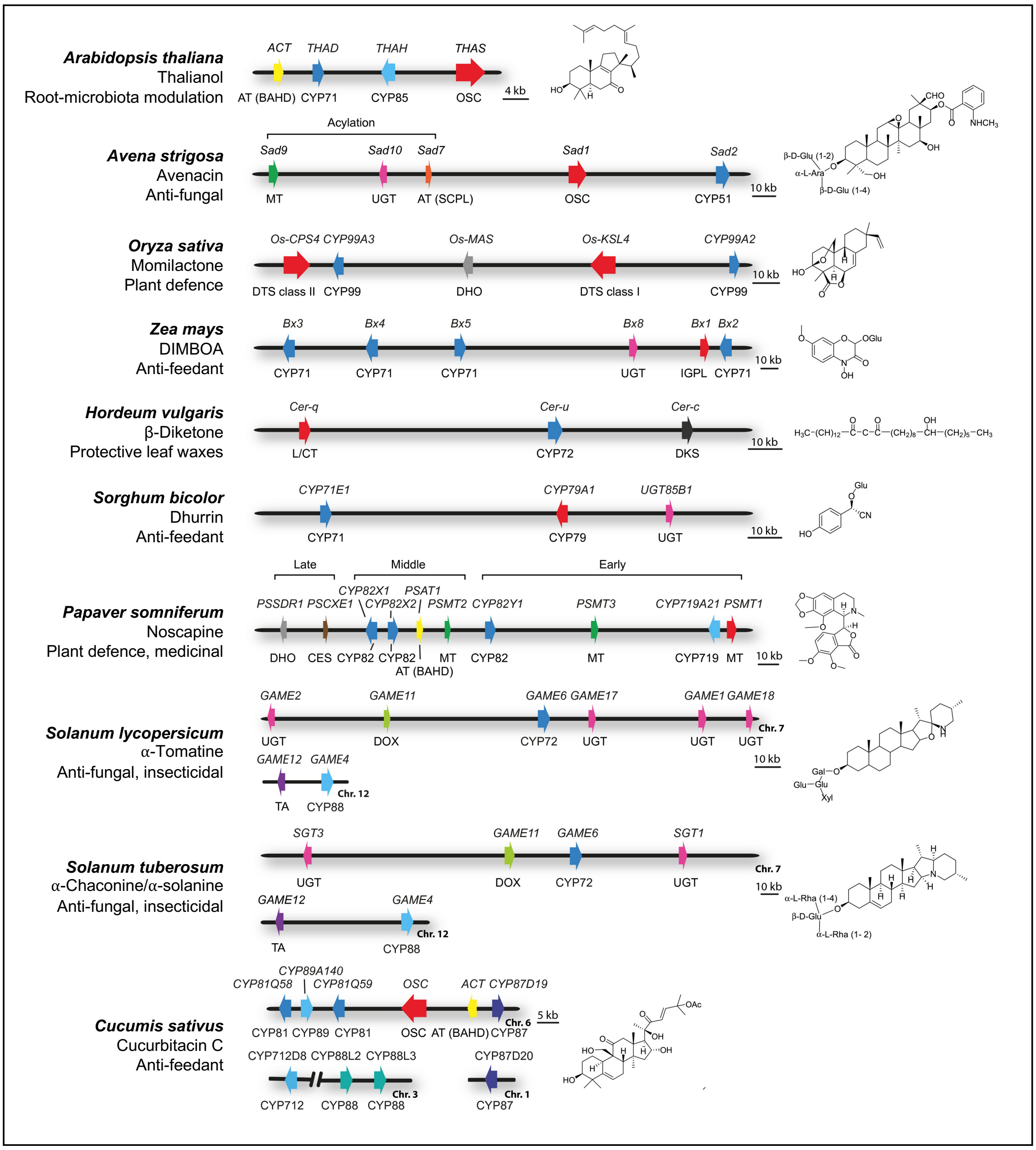


Figure 1.2 **Plant biosynthetic gene clusters**

Examples of various BGCs from different plant species are shown, along with their *in planta* roles. The gene(s) for the first committed pathway step are indicated in red. A range of clustering types and natural products classes are shown. Adapted from [32].

A great deal is unknown about scope, regulation and evolution of plant BGCs. Certainly, not all pathway genes for plant specialised metabolites are found in BGCs. There are a number of hypotheses as to how and why BGCs occur in plants.

A likely origin of genes for specialised metabolism comes from gene duplication and neofunctionalization from primary metabolism [28], as relaxed selection pressures allow the evolution of novel chemistries. Recent studies in the Brassicaceae have demonstrated that recruitment of genes to a specific locus appears to be highly dynamic, where superficially homologous BGCs in related species have been shown to be derived from independent origins [35]. Across the eudicots, terpene synthases and cytochrome P450s have been observed to act as ‘microsyntenic’ gene blocks [36], and miniature inverted transposable elements (MITEs) have been implicated in BGC formation and regulation [37].

The presence of BGCs may be selected for due to the potential for specialised metabolic pathways to create toxic intermediates, therefore tight co-regulation is needed [33,38]. It has been argued that co-localisation prevents the loss of key pathway genes during recombination events [33]. Furthermore, the local chromatin environment may provide a particular means for gene expression to be tightly controlled. In *A. thaliana*, chromatin marks have been observed to be strongly associated with repression and expression of BGCs [32,39].

1.4 Triterpenes

Genes encoding for triterpene biosynthesis are found in BGCs across monocots and dicots [32], and genome analyses have shown co-located *OSC-CYP* gene pairs are distributed non-randomly throughout plant genomes [36]. Triterpene BGCs also provided the basis for fundamental studies in the Brassicaceae demonstrating the remarkable ability of plants to independently assemble BGCs from ancestral gene blocks [35].

Triterpene are C30 terpenoids, the largest class of plant specialised metabolites, and have a wide variety of roles *in planta*. Sterols are triterpenoids essential for the controlling cell membrane fluidity, and the large family of steroid signalling hormones are derived from them [40]. As plant specialised metabolites, common role for triterpenes is as part of plant defence, such as the production of waxy cuticle layers, the protection against feeding by the production of insecticidal or bitter-tasting compounds and defence against soil-borne pathogens by anti-fungal compounds [16,33,41–43]. A recent study of a complex triterpene metabolic network in *Arabidopsis thaliana* has demonstrated how a range of molecules are used to modulate population of soil microbiota [1]. Furthermore, triterpenes have also been implicated in growth and developmental pathways [38,44,45].

For humans, triterpenes have found a wide range of uses. Medicinally, triterpenes are reported to exhibit a wide range of activities, including as anti-inflammatories, neuroprotectives, antivirals, cytotoxic and cytoprotective agents [46–52]. Perhaps the most prominent triterpene used in a medicinal context is the vaccine adjuvant QS-21 isolated from *Quillaja saponaria* [53]. Outside of medicine, triterpenes are used as foaming agents, insecticides, fungicides, piscicides, soaps and sweeteners [12,43,51,54].

The first committed step of triterpene biosynthesis is the cyclisation of 2,3-oxidosqualene, derived from the mevalonate (MVA) pathway. This is catalysed by a family of enzymes known as oxidosqualene cyclases (OSCs) or simply ‘triterpene synthases’. This results in the production of a triterpene ‘scaffold’, which is then functionalised by cytochrome P450s (CYPs) via oxidation at specific C positions. Tailoring enzymes, such as glycosyl-transferases (GTs), methyltransferases (MTs) and acyltransferases (ATs) are then able to act at these positions to modify the scaffold [41]. This is summarised in Figure 1.3. This process results in a huge array of triterpenes all derived from a single precursor, with over 100 triterpene scaffolds and over 20,000 triterpene compounds having been isolated from nature [21,41].

In terms of our ability to predict and test enzyme function, the *Nicotiana benthamiana* transient expression system has proven highly effective at rapidly screening candidate enzyme activity [55]. It has also has allowed the rapid production of triterpene specialised metabolites at a gram-scale [21]. Therefore, in developing a high-throughput, systematic synthetic approach for the study of plant BGCs, triterpenes stand out as ideal candidates for further investigation due to their variety of biological activities, propensity to form BGCs, tractable biosynthetic pathways and the existence of proven screening and characterisation platforms for candidate genes.

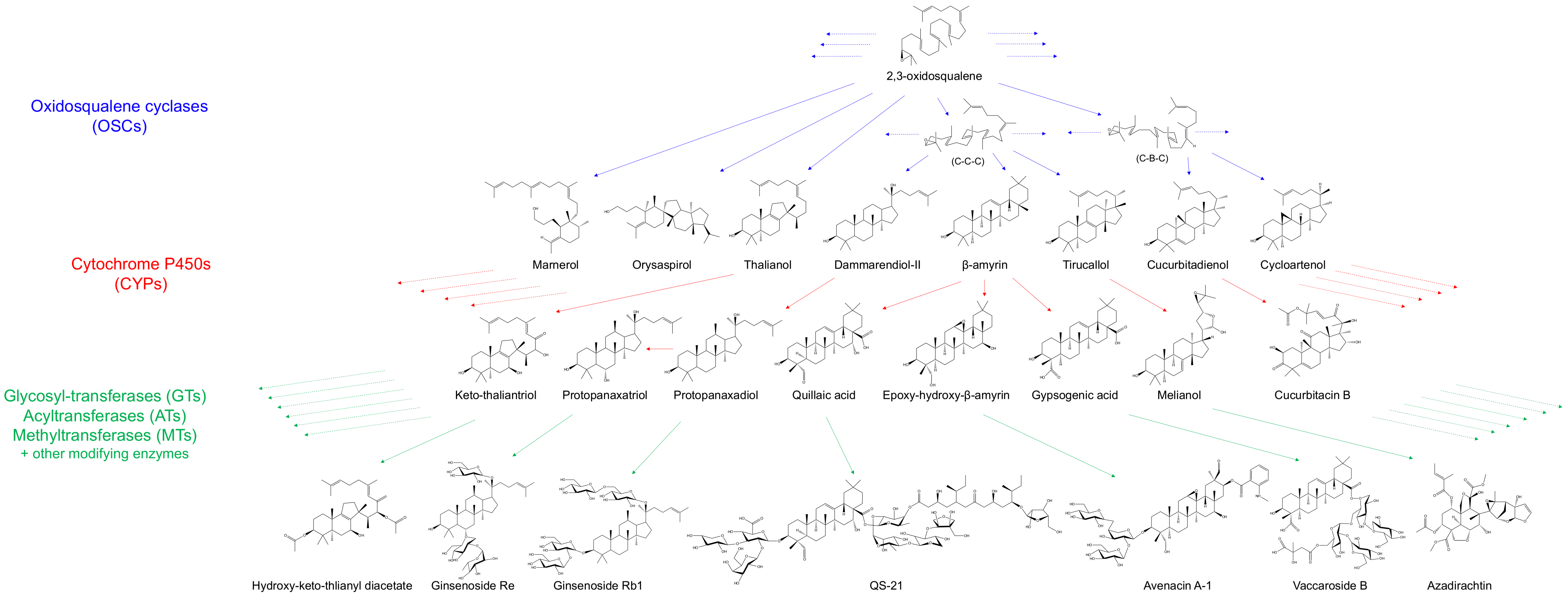


Figure 1.3 Examples of triterpene biosynthetic pathways

Triterpene diversity is generated from the single substrate 2,3-oxidosqualene, which is cyclised into a triterpene scaffold by OSCs. These scaffolds are oxidised at specific carbon positions by CYPs, and this functionalisation allows modifying enzymes such as GTs, ATs and MTs to further decorate the scaffold. The enormous variety of triterpenoid compounds isolated from nature can be broadly be assigned to the activities of these key enzyme groups.

1.5 Thesis summary

The aims of this PhD are as follows: to perform broad, systematic BGC mining of available plant genomes using available tools (Chapter 2); to perform a comprehensive bioinformatic analysis of OSCs across all available plant genome data (Chapter 3); to investigate the reported phenomenon of OSC-CYP co-evolution and co-localisation (Chapter 4); to build tools for the prediction of GT function (Chapter 5); to comprehensively report on the wider nature of triterpene biosynthetic enzyme co-localisation (Chapter 6); and, to investigate specific plants and BGCs and to provide case-studies into the nature of triterpene biosynthetic genetic organisation in *Quillaja saponaria* (Chapter 4) and *Avena strigosa* (Chapter 7). To achieve this in a high-throughput and systematic manner which aligns with the ethos of rational design within synthetic biology, multiple bioinformatic and computational tools have been required to be built or sourced, tested and optimised. The development of such tools and approaches consequently forms an integral part in achieving the aims of this project. The outputs of this project therefore are to build a broad understanding of plant BGC prevalence and characteristics, and to use triterpene biosynthetic enzymes as an example to investigate this deeply. Evidently, there are numerous opportunities throughout this process to leverage the data for *in silico* prediction of biosynthetic activity, which, in conjunction with collaborators, can be tested.

Chapter 2. Application of plant genome mining tools

2.1 Introduction

Given the discovery of plant BGCs and their potential for streamlining pathway discovery methods, the development of plant genome mining tools has been a recent research focus [26,32,56–59]. The aims of these are broadly to provide systematic analyses of submitted genome sequence data and subsequently report putative BGCs, potentially with some information as the predicted functions of the constituent genes. Such tools are a necessary part of developing a coherent synthetic biology approach to plant metabolic science, as well as a potentially important method for determining the scale and scope of BGC prevalence amongst plants [Chapter 1].

Three recently developed tools are ‘plantiSMASH’ [34], ‘PhytoClust’ [58] and ‘PlantClusterFinder’ [59]. Figure 2.1 shows a summary of their methodology. plantiSMASH and PhytoClust are both built using the framework of antiSMASH (a tool developed for the discovery and analysis of microbial and fungal BGCs [60]) and so share a similar approach, whereas PlantClusterFinder is part of the broader ‘Plant Metabolic Network’ gene and pathway classification pipeline [59,61]. For the purposes of this text, ‘PlantClusterFinder’ will refer to this whole pipeline, as summarised in Figure 2.1.

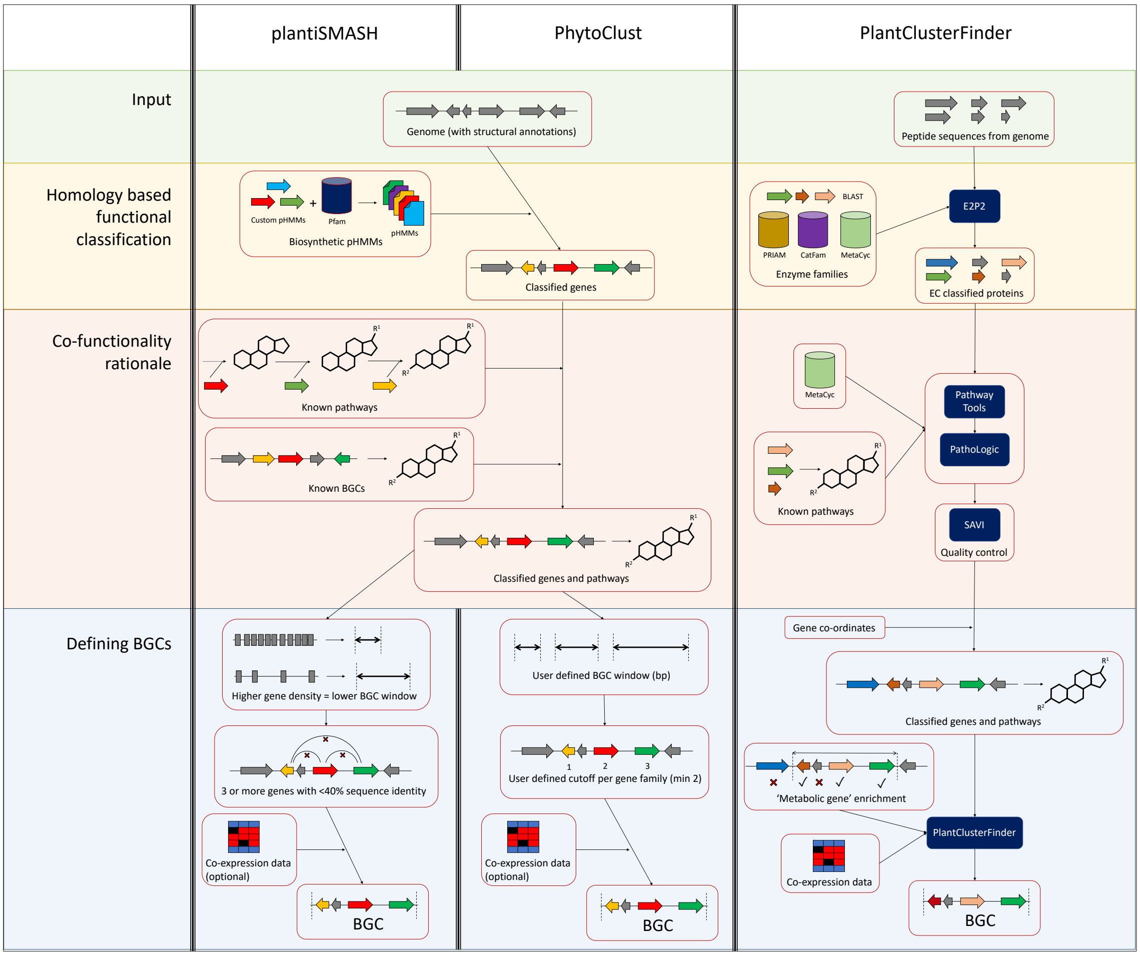

plantiSMASH
PhytoClust
PlantClusterFinder
Input
Genome (with structural annotations)
Peptide sequences from genome
Homology based functional classification
Custom pHMMs
Pfam
pHMMs
Biosynthetic pHMMs
Classified genes
BLAST
PRIAM
CatFam
MetaCyc
Enzyme families
E2P2
EC classified proteins
Co-functionality rationale
Known pathways
Known BGCs
Classified genes and pathways
MetaCyc
Pathway Tools
PathoLogic
Known pathways
SAVI
Quality control
Defining BGCs
Higher gene density = lower BGC window
3 or more genes with <40% sequence identity
Co-expression data (optional)
BGC
User defined BGC window (bp)
User defined cutoff per gene family (min 2)
Co-expression data (optional)
BGC
Gene co-ordinates
Classified genes and pathways
'Metabolic gene' enrichment
PlantClusterFinder
Co-expression data
BGC

Figure 2.1. **Simplified graphical summary of three plant BGC mining and annotation tools** The three tools recently developed to mine plant genomes for BGCs are plantiSMASH [34], PhytoClust [58] and PlantClusterFinder [59]. plantiSMASH and PhytoClust are built on the same framework, and so share many attributes. All of these approaches use homology-based classification, although PlantClusterFinder derives this from putative enzymatic activity instead of alignment to via pHMMs. The methodology for defining BGCs is different across all three approaches.

A genome with structural annotations (i.e. putative gene models) is the required input for plantiSMASH and PhytoClust. This is due to their reliance on HMMer [62] to first characterise biosynthetic enzymes, which uses protein sequence data. The classifications in both tools are built from profile Hidden Markov Models (pHMMs) within the Pfam database [63], where known biosynthetic families are suitably represented, as well as custom pHMMs derived from characterised plant biosynthetic proteins. The result of this is that only the targeted gene families are subsequently classified. PlantClusterFinder instead classifies all of the protein sequences from a candidate genome using Enzyme Commission (EC) classifications via homology to a comprehensive set of known enzymes, resulting in a much larger relative set of classified sequences [59].

In plantiSMASH and PhytoClust, the rationale for a set of characterised enzymes being putatively co-functional is determined by 'cluster definitions', which are lists of gene families known or presumed to act in a shared specialised metabolic pathway [34,58]. These are customisable, but in their default state are based on known BGCs and biosynthetic pathways. These tools therefore target a specific metabolic space. In plantiSMASH these are reported with generic descriptors such as 'terpene', 'alkaloid' and 'saccharide'. Conversely, PlantClusterFinder takes a large-scale approach by creating a global metabolic model for the protein set in question. This process is guided by known metabolic reactions and, due to its complexity, undergoes a specific quality control and validation pipeline [59] before genes are mapped onto physical genomic space.

The results of these approaches are various set of genes predicted to be co-functional in some metabolic pathway of interest. The definition of what precisely constitutes a BGC is non-trivial, and made challenging by the highly variable nature of plant genome structure (e.g. gene density, intron size, genome size, ploidy etc) as well as the limitations in the classification methods used to differentiate functionally divergent genes [26,64,65].

PlantClusterFinder assesses the enrichment of genes involved in 'specialised metabolism' (as defined by EC denoted pathways) across putative BGCs. It determines the 'ends' of the BGC by a significant co-location of genes, which are predicted to act in a shared pathway, in comparison to the distribution across the whole genome. Given that this returns potentially thousands of putative BGCs, co-expression data is used to select the best BGC candidates [59].

PhytoClust requires the user to determine both the maximum and minimum BGC sizes (in bp) as well as the minimum number of separate gene families required to report a BGC [58]. This allows a large degree of customisability but means that some optimisation is likely required for each species of interest analysed. Furthermore, whilst the pHMMs used are designed to represent functionally distinct gene groups, the resolution to which they resolve alternatively functioning enzymes within the same gene superfamily is variable. For example, two cytochrome P450s (CYPs) which are divergent in both sequence and function will be classified as the same gene family using Pfam definitions.

To solve the issue of variable plant genome structure, plantiSMASH uses a dynamic algorithm which accounts for global (i.e. across the whole genome) and local (i.e. within the region of interest) gene density [34]. This is valuable, because gene density often changes dramatically across a chromosome [34,64], so even a static BGC definition may be unsuitable for retrieving the whole BGC complement. To define the number of gene family co-located within the region determined by this algorithm without inheriting the biases of the pHMMs used, a 40% sequence identity cut-off is utilised. As such, plantiSMASH defines a BGC as a minimum of three co-located genes all of which share no more than 40% sequence identity with each other [34].

Whilst PlantClusterFinder is comprehensive, its dependence on a complex annotation and database generation pipeline means it is broadly unsuitable for running locally and high-throughput screening of genome data as it becomes available. Furthermore, EC classification is generally unsuitable for detailed analysis of triterpene biosynthetic pathways, as it is generally not amenable to capture the evolutionary relationships between enzymes with convergently evolved functions. Both plantiSMASH and PhytoClust are much more suitable for this, and, given that plantiSMASH handles variation in gene density automatically, plantiSMASH is the tool which will be used herein.

2.1.1 Aims

There has not been an investigation into how such approaches specifically handle triterpene biosynthetic enzymes, beyond proof that the known triterpene BGCs in monocots and dicots are returned. The aims of this chapter are therefore to assess the current 'baseline' using plantiSMASH 1.0 and determine where, if needed, changes to this approach need to be made for a comprehensive survey of plant triterpene BGCs. This chapter's aims are therefore to:

- Collate a set of suitable plant genomes and analyse them with plantiSMASH 1.0
- Investigate the reporting of triterpene biosynthetic genes in terms of accuracy of annotation and capability for functional prediction

2.2 Methods

2.2.1 Genome collection

595 publicly available Viridiplantae genomes were sourced from the NCBI genome database (www.ncbi.nlm.nih.gov/genome/), Phytozome v11 (phytozome.jgi.doe.gov/), CoGe (genomevolution.org/) and other individual sequencing repositories. Summary data for these genomes are given in Table A1.

2.2.2 plantiSMASH and OSC counting

plantiSMASH 1.0 and its dependencies were installed according to the developer's instructions [34]. Suitable genomes (i.e. those with structural annotations comprising gene models and putative protein sequences) were put forward for BGC mining by plantiSMASH 1.0. Standard parameters were used, other than removing the default maximum analysis limit of 9999 contigs. HTML and JavaScript outputs were parsed by Python. As in plantiSMASH, OSCs across the whole genome were defined by alignment to the Pfam profiles 'SQHop_C' (PF13243) or 'SQHop_N' (PF13249), with the highest scoring sequence taken forward where multiple isoforms were present in the annotation.

2.3 Results

2.3.1 A wide range of terpene BGCs across plants are found using plantiSMASH 1.0

A total of 273 genomes, representing 177 Viridiplantae species, were analysed using plantiSMASH 1.0 [34]. This returned a total of 9350 putative BGCs of which 1866 were classified as ‘terpene’, meaning they contained at least one putative terpene synthase. Figure 2.2 demonstrates the variability of putative BGC distribution and class across plant clades and genomes. Green algae (Figure 2.2A) are reported to contain few to no BGCs, whereas monocots (Figure 2.2B), Kalanchoe and Caryophyllales (Figure 2.2C) and Brassicales (Figure 2.2D) all return a range of BGC counts. It must be noted that these genomes are variable in their assembly quality (Table A1), therefore certain genomes are likely to have their BGC counts underreported due to genome fragmentation.

Whilst OSCs are an evolutionarily distinct family compared to other terpene synthase enzymes [36,66], plantiSMASH does not differentiate in their BGC classification. Therefore, to assess the presence of putative triterpene BGCs, manually screening of the reported data for the Pfam profiles ‘SQHop_C’ (PF13243), ‘SQHop_N’ (PF13249) and ‘Prenyltrans’ (PF00432) (all of which correspond to OSC sequences) is required. After this screening, 348 of the 1866 ‘terpene’ BGCs were found to contain OSCs.

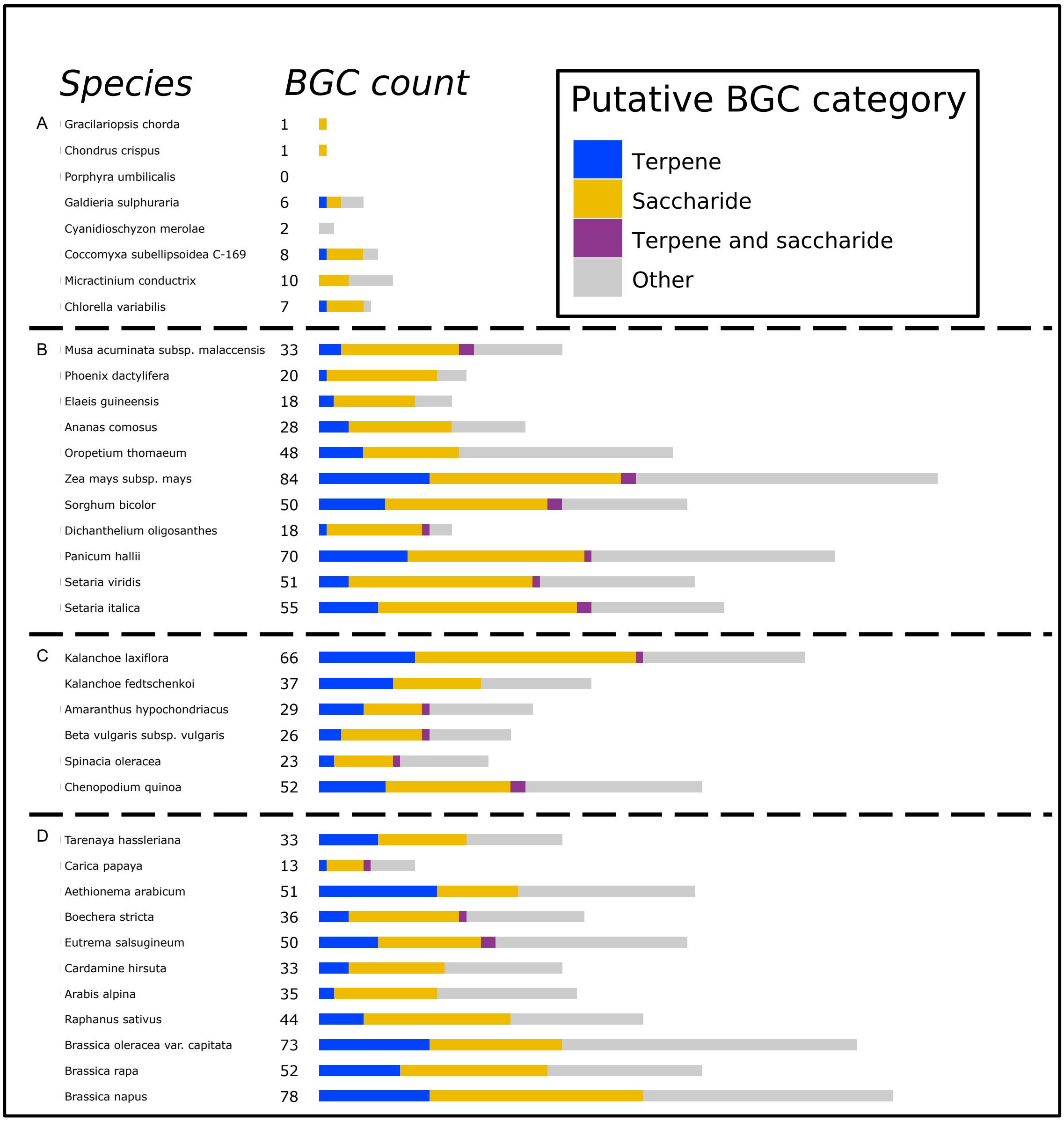


Figure 2.2. **Putative plantiSMASH 1.0 BGC counts and classifications for example species** Length of stacked bar charts represent the total BGC count from each species. BGC categories, as defined by plantiSMASH 1.0, are represented by the colours shown in the key. A) Green algae B) Monocots C) Saxifragales and Caryophyllales D) Brassicales

Figure 2.3 shows some examples of putative, uncharacterised triterpene BGCs reported by plantiSMASH 1.0. The variation in BGC size is evident, as is the presence of intervening, non-biosynthetic genes (grey). It is clear that this tool is able to locate enzymes of interest to motivate further study. However, no information is available as to a more specific functional classification. It is known that there are both sequence-function relationships and distinct phylogenetic clades for classification of many triterpene biosynthetic enzymes [30,41], therefore some further detail beyond categorisation as a ‘terpene’ BGC should be possible.

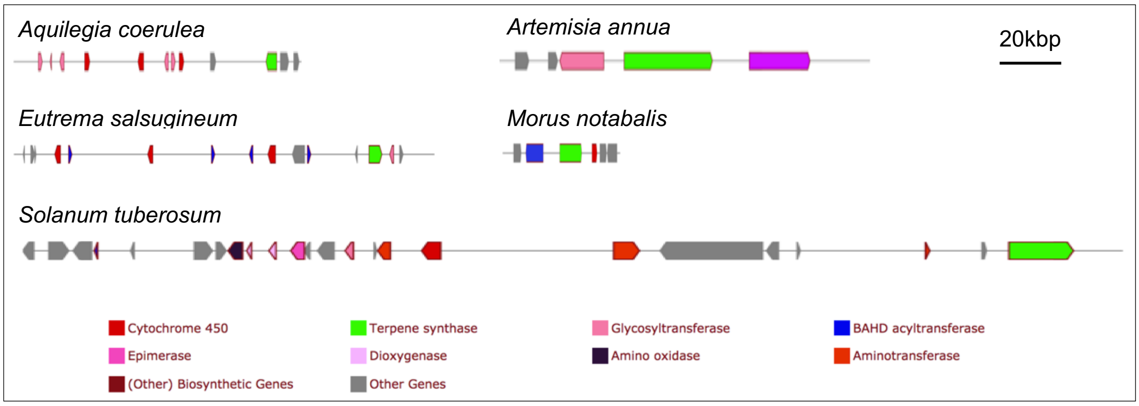


Figure 2.3 **Example putative triterpene BGCs discovered by plantiSMASH 1.0**
Putative BGC genes coloured according to key and scaled to demonstrate variation in triterpene BGC component gene families as well as BGC size and gene density. Absolute BGC size and density is related to the overall plant genome size as well as chromosomal location.

These data also provide an opportunity to assess triterpene BGC occurrence across plant species. Figure 2.4 shows a taxonomy of 47 plant species for which full chromosome level assemblies were available (Table A1). The bar charts display the total number of putative OSCs found in each species, as well as whether they were found to be part of a putative BGC. These data show that the proportion of OSCs found in BGCs, according to plantiSMASH 1.0, is relatively variable across plant species. For example, *Solanum lycopersicum* and *S. pennellii* appear to have the majority of their OSCs 'clustered', whereas the Malpighiales show the inverse.

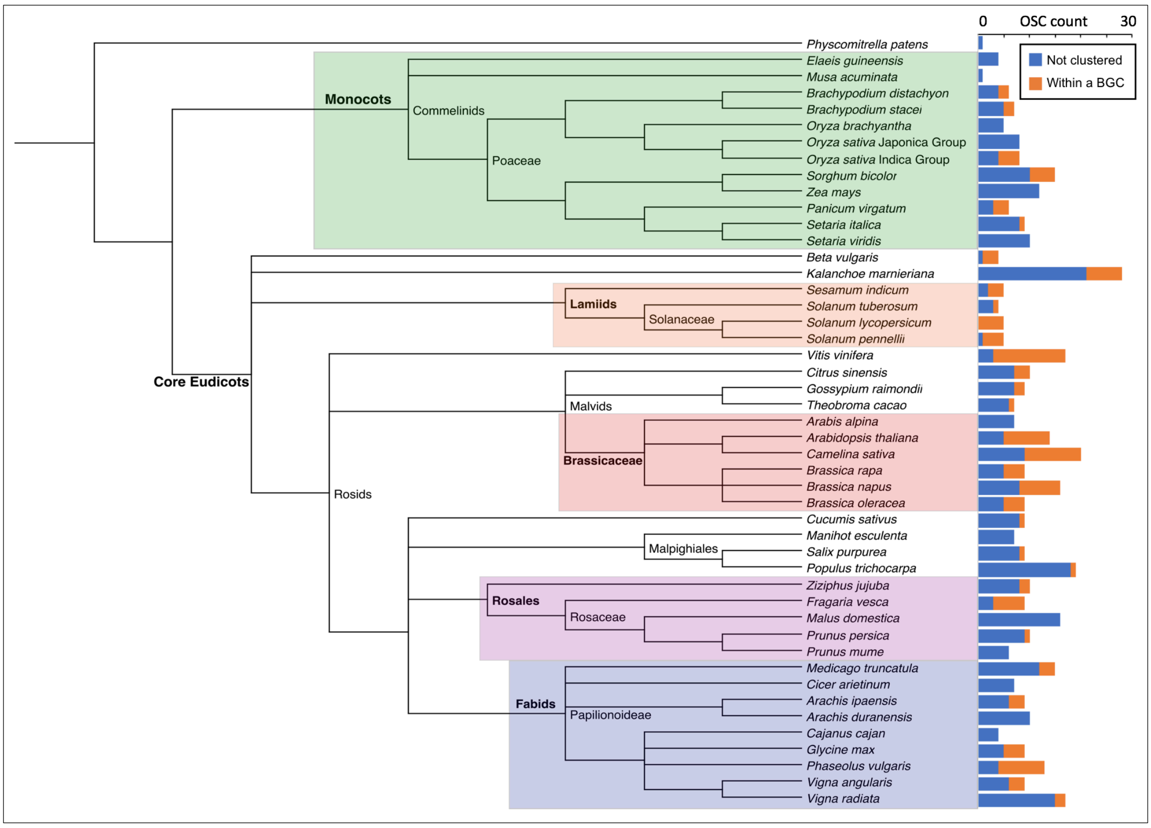


Figure 2.4 **Proportion of *OSCs* that form part of putative BGCs**

Plant species for which a chromosome-level assembly was present were analysed by plantiSMASH. The counts of *OSCs* in the genome and those of which were assigned to putative BGCs are shown.

2.3.2 Output accuracy depends on the variable annotation quality of input genomes

Whilst the data presented above may appear promising, upon closer inspection of putative BGCs it is apparent that the quality of the genome's structural annotations are fundamental in determining mining accuracy. Specifically, plantiSMASH does not use any filtering for the quality of pHMM alignments beyond the defaults of HMMer [34]. This is partly because of the relatively low availability of well-characterised BGCs and specialised metabolic pathways available during the development of plantiSMASH. Without this generalisation, the scope of plantiSMASH would be quite limited.

Nonetheless this can lead to undesirable consequences. First, low-quality, pseudogenic and/or truncated protein sequences are often present in putative BGCs. Furthermore, because the quality of plant genome data is often highly variable (Chapter 1), it raises questions as to the comparability of data. Finally, because genes involved in plant specialised metabolism are often expressed in very specific conditions and/or tissues [26,32], it is possible that these genes families will disproportionately suffer from missing annotations. To demonstrate this, the representative genomes at the beginning of this project for *Oryza sativa* Japonica Group (GCA_001433935.1) and *Oryza sativa* Indica Group (GCA_000004655.2) respectively contained eight and three of the 12 manually annotated OSC sequences (as described in [66]).

2.4 Conclusions

High-throughput, systematic methods for mining plant genomic data are in their infancy, primarily because data have only recently been generated in sufficient quantity to warrant such approaches. The tools described in this chapter demonstrate the challenges in working with plant genomes and the opportunities these provide for innovation and novel bioinformatic approaches. However, by utilising triterpene biosynthetic enzymes as a model for BGC mining, it has been shown that there is scope for improvement and refinement of these methodologies. Without underpinning data of reasonable quality, bioinformatic analyses will be unable to answer fundamental questions about the evolution and diversification of such genes, nor will they be able to accurately predict enzyme function and activity.

Chapter 3. Comprehensive genome mining for OSCs

3.1 Introduction

3.1.1 OSCs in plants

As discussed in Chapter 1, OSCs catalyse the first committed step of the triterpene biosynthetic pathway via the cyclisation of 2,3-oxidosqualene into a triterpene backbone. These scaffolds are diverse, ranging from monocyclic to pentacyclic structures. For most penta- and tetra-cyclic triterpenes, conformational arrangement via the dammarenyl or protosteryl intermediate cations separates these compounds into the 'sterols' and the 'triterpenes' respectively [41]. Whilst the scope of this thesis deals with triterpene biosynthesis, there is natural overlap with sterol biosynthetic enzymes, as well as edge cases and alternate biosynthetic pathways [67], so these will also be studied here.

Previous phylogenetic work has demonstrated that OSCs show some degree of sequence-function relationship, in that certain phylogenetic clades of OSC sequences have shared function across a wide range of plant species [41,66]. Furthermore, various plant taxonomic groups can have specific repertoires of OSC subtypes and the evolution of OSCs across the Viridiplantae appears to show convergent evolution of shared function across evolutionary divergent sequences [41]. As such, whilst the sequence-function relationship is not as disordered as e.g. sesqui- or di-terpene synthases [6], there is still scope for complex relationships between sequence, structure, function and the evolutionary pressures that guide them.

Functional characterisation and mutagenesis of OSCs demonstrates the dynamic potential for rapid diversification of enzyme activity. For example, two very closely related OSCs in rice produce highly distinct chemical compounds orysatinol and parkeol. For each enzyme, the mutation of three amino acids is sufficient to convert functionality from one to the other [67]. In another case, single amino acid changes are able to modulate product specificity in SAD1, an OSC from *Avena strigosa*, and LUP1, from *Arabidopsis thaliana* [34]. However, there has not been a more generic success in determining a universal sequence-structure-function relationship model, with most studies utilising substrate docking and analysis for rationalisation of specific reactions of interest [67–69].

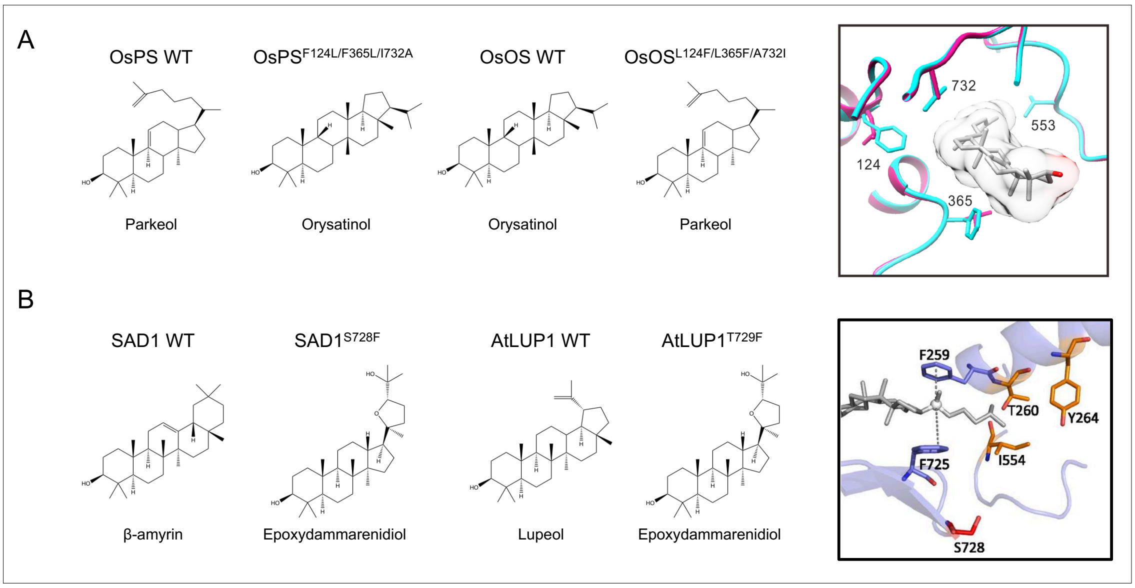


Figure 3.1 **Key OSC residues discovered to confer functional specificity**
Mutagenic studies have identified residues in determining OSC activity. A) Three residues are able to modulate production between two contrasting biosynthetic products in OSCs isolated from *Oryza sativa* B) Mutation of a single conserved residue in functionally distinct OSCs from *Avena* strigosa and *Arabidopsis thaliana* results in the production of epxoydammarendiol. Homology models adapted from [67] (A) and [69] (B) each showing the OSC catalytic site.

3.1.2 Overcoming variable genome quality

In order to access plant genomes with poor or no structural annotations, a solution is required which can rapidly and accurately generate annotations based only on DNA sequence and homologous protein sequences of the families of interest. Numerous tools exist to achieve this, which can broadly be split into '*ab initio*' and targeted approaches. The former relies on generic models/rules of global gene occurrence and requires training on annotated genome data to learn these. These approaches generally produce a large number of gene annotations across a given genome, as they are built to predict all target genes. Examples of such *ab initio* tools are Augustus [70] and GlimmerHMM [71] (which is included in plantiSMASH 1.0 for optional gene prediction [34]).

Conversely, targeted approaches rely on the input of sequences and/or alignments of the gene family of interest. These tools can range in complexity and computational scale, from joining BLAST high-scoring pairs into a coherent gene model [72], to global, exhaustive protein-to-genome alignment algorithms [73,74]. Examples of these tools are Augustus-PPX [70], GenBlastG [72], Exonerate [73] and Selenoprofiles [74].

3.1.3 Aims

The aims of this chapter are to:

- Trial various gene prediction tools in order to find an approach that gives accurate results and can be utilised in a systematic mining pipeline
- Utilise this in order to extract all putative OSCs from all genome sequence data available
- Perform phylogenetic analysis on the OSCs to observe their evolutionary diversity across the Viridiplantae
- Investigate OSC sequence-function relationships

3.2 Methods

3.2.1 Testing alternate annotation approaches

Augustus/Augustus-PPX [70], Exonerate [73], GlimmerHMM [71] and Selenoprofiles [74] were tested for target gene annotation. Augustus and GlimmerHMM are *ab initio* methods which were run with default settings using the trained plant models included with the packages. Augustus-PPX, Exonerate and Selenoprofiles are profile-based methods and therefore required the generation of alignments of the monophyletic gene families of interest. Selenoprofiles is a multi-step pipeline that includes the use of Exonerate as part of the annotation process (Figure 3.2). For these tools, profiles were derived from the characterised OSC and CYP sequences described in [41].

The *A. thaliana* and *O. sativa* Japonica Group genomes were used to test the above tools, with the aim to regenerate the true annotations of OSC and CYP sequences in these genomes. For profile-based annotation, the sequences derived from *Arabidopsis* species and *Oryza* species were removed from the alignments. For *ab initio* methods, all pre-packaged plant models were tested and the most accurate used for comparison (despite this being a 'best-case' scenario, particularly given unannotated genomes of interest are unlikely to be closely related to model species).

Optimisation was carried out in Selenoprofiles as it was not designed for plant genome data as default, where intron sizes can be well over 10Mbp and alignment scores of candidate genes to the closest known profile can be relatively poor. The same parameters used for the Exonerate stage of the Selenoprofiles pipeline were used for testing Exonerate as a standalone. The non-standard parameters used for were as follows:

```
p2g_filtering = len(x.protein()) >40 or x.coverage()> 0.3
p2g_refiltering = x.awsi_filter(awsi=0.2)
exonerate_opt = --score 300 --maxintron 20000
genewise_opt = -splice flat
blast_filtering = x.evalue < 1e-5 or x.sec_is_aligned()
```

This mining approach was applied to all plant genomes available, including those with structural annotations, in order to maximise yield. Wherever prior annotations overlapped with the putative annotations generated here, the prior annotations were always selected.

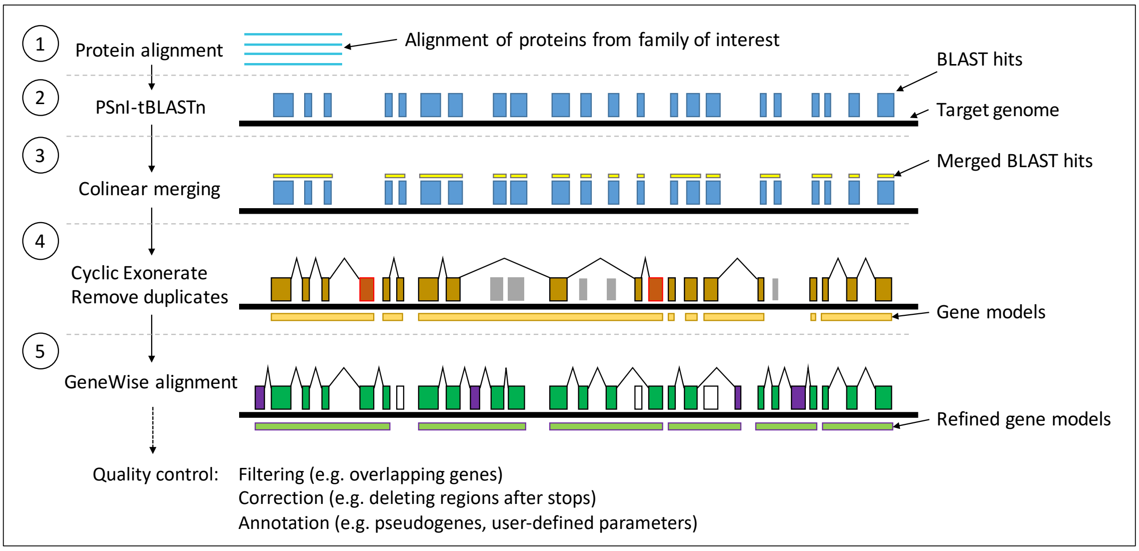


Figure 3.2 **Summary of computational workflow for Selenoprofiles**
Summary of the Selenoprofiles pipeline showing the key alignment steps. The input consists of an alignment of protein sequences from the family of interest and a nucleotide genome sequence. Iterative tBLASTn is used to generate initial homology blocks, which are then merged according to co-linearity with the profile sequences. Exonerate and GeneWise are then used to refine the protein to genome alignments around these regions, before a final filtering step to remove overlaps and flag pseudogenes.

3.2.2 OSC mining and phylogenetics

Genome mining for OSCs using Selenoprofiles and HMMer was carried out on 304 plant genomes representing 258 Viridiplantae species as described in Chapter 2, using a profile generated from an alignment of the 82 characterised OSC sequences described in [41]. For Selenoprofiles, parameters were as described above. For HMMer, a bitscore cutoff of 500 was used to select putative OSC annotations, which was derived via manual inspection of outputs from well-characterised genomes. One genome per species was chosen for subsequent analysis based on the number and quality of putative enzymes found. This resulted in the generation of 2068 unique, non-overlapping putative OSC sequences.

Before alignment, high-quality putative protein sequences were filtered by requiring a minimum length of 650 amino acids and the removal of pseudogenes as flagged by Selenoprofiles (i.e. if frameshifts or indels were required to generate the protein profile to nucleotide alignment). This produced 1404 high-quality, full-length putative OSC sequences, which were aligned with 82 characterised OSCs described in [41]. Alignments were carried out with MAFFT [75] using the global pairwise alignment model. A phylogenetic tree was generated with RaXML [76] using automatic model selection with the gamma model of rate heterogeneity with 100 runs and bootstraps. Tree topology was subsequently confirmed via MrBayes [77]. A summary of this methodology is given in Figure 3.3.

3.2.3 Profile generation

pHMMs of representative sequences within each phylogenetic OSC group were generated by selecting up to 100 representative samples across each clade followed by aligning and building with HMMer [62]. These profiles were then used for on-the-fly characterisation of OSC sequences by choosing the profile that most closely matched the OSC in question. An alignment score cut-off was not used, but instead filtering was achieved via a minimum alignment span of 450 amino acids, which was found to maintain accuracy whilst allowing putative classification of sequences not included in the phylogenetic analysis. It is noted that not all groups are monophyletic, so accuracy is reduced when attempting to assign proteins to specific groups based on sequence similarity alone for these groups.

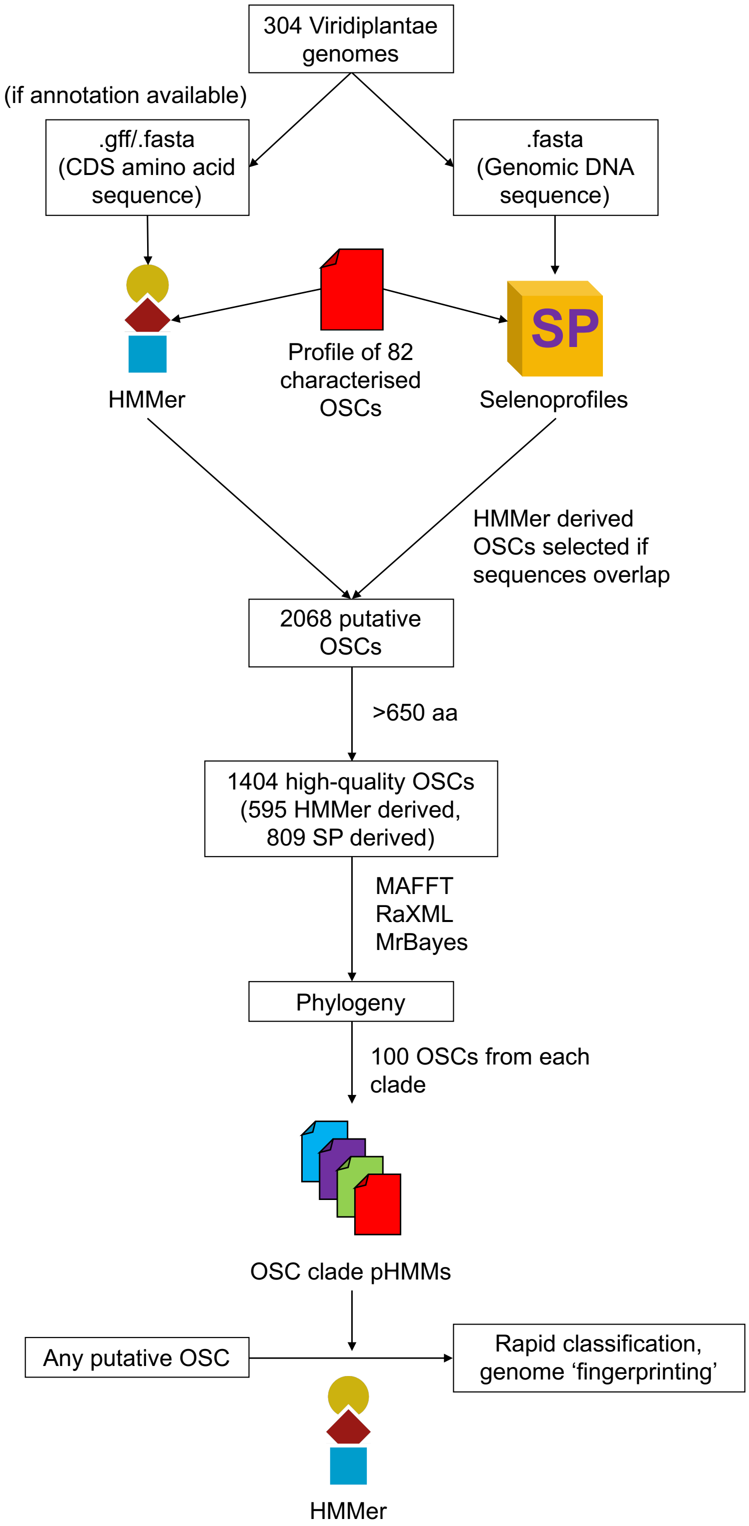

304 Viridiplantae genomes
(if annotation available)
.gff/.fasta (CDS amino acid sequence)
.fasta (Genomic DNA sequence)
SP
HMMer
Profile of 82 characterised OSCs
Selenoprofiles
HMMer derived OSCs selected if sequences overlap
2068 putative OSCs
>650 aa
1404 high-quality OSCs (595 HMMer derived, 809 SP derived)
MAFFT
RaXML
MrBayes
Phylogeny
100 OSCs from each clade
OSC clade pHMMs
Any putative OSC
Rapid classification, genome 'fingerprinting'
HMMer

Figure 3.3 **Summary of computational workflow for high-throughput, systematic OSC mining from plant genomes of varying annotation quality**

HMMer and Selenoprofiles were used in conjunction with characterised OSCs in order to fully utilise the available plant genome sequence data, despite the absence or quality of genome annotations. After filtering the discovered OSCs to ensure only high-quality sequences were assessed, a phylogenetic analysis was carried out, which was then used to define distinct OSC groups and pHMM generation for on-the-fly OSC characterisation.

3.3 Results

3.3.1 Selenoprofiles is the most suitable tool for extracting putative proteins from unannotated genome data

The tools used in trialling methods for rapidly and accurately extracting putative biosynthetic genes were: Augustus/Augustus-PPX [70], Exonerate [73], GlimmerHMM [71] and Selenoprofiles [74]. These all vary significantly in methodology, implementation and results. To summarise, profile-based methods are most accurate for finding specific gene families whereas *ab initio* methods return a genome-scale complement of putative genes [70,73].

Of the profile-based tools tested here, Selenoprofiles was by far the most accurate in terms of protein sequence identity (Figure 3.4). Selenoprofiles also proved amongst the easiest to implement. Because this tool was not designed for plant genomic data, it required optimisation to ensure that sufficiently large intron sizes were allowed for, as well as a more lenient alignment score filter for putative gene assignment to a given profile (see 3.2 Methods for details). After this, it was able to find OSC and CYP sequences in *O. sativa* var. Japonca and Indica genomes with an average protein sequence identity of 98% in comparison to the true sequences. This is due to its comprehensive, multi-step pipeline where multiple alignment tools are applied a sequential manner [74] (Figure 3.2).

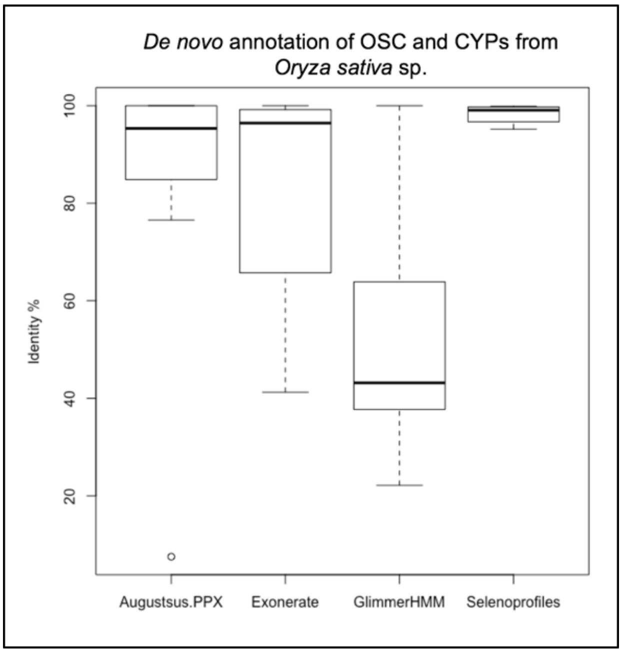


Figure 3.4 **Testing gene finding tools to extract putative protein sequences from example unannotated genome data**
Prediction accuracy of various tools to annotate OSC and CYP protein models in the *O. sativa* var. Japonca and Indica genomes. Data for *ab initio* Augustus annotation is not shown, but it performed considerably worse than Augustus-PPX. Identity score represents the sequence identity between the predicted and known OSC amino acid sequences. Default plotting parameters in R are used, with the height of the box covering the interquartile range (IQR), and the whiskers range using a value of 1.5x the IQR.

3.3.2 Mining and phylogenetics of OSCs across the Viridiplantae shows sequence-function relationship and clade specific diversification

Selenoprofiles based mining was carried out on 304 plant genomes representing 258 species within the Viridiplantae (Table A1). Putative OSC genes were obtained from these genomes, numbering 2068 unique, non-overlapping sequences of which 1404 were high quality. Of these, 809 OSC sequences were derived from unannotated genome data using Selenoprofiles as described above. For comparison, Xue et al. [66] assessed 96 OSCs from 16 species. A maximum-likelihood tree of these 1404 sequences plus 82 characterised OSCs [41] is shown in Figure 3.5. Letters are used to denote the various OSC clades, which are discussed below.

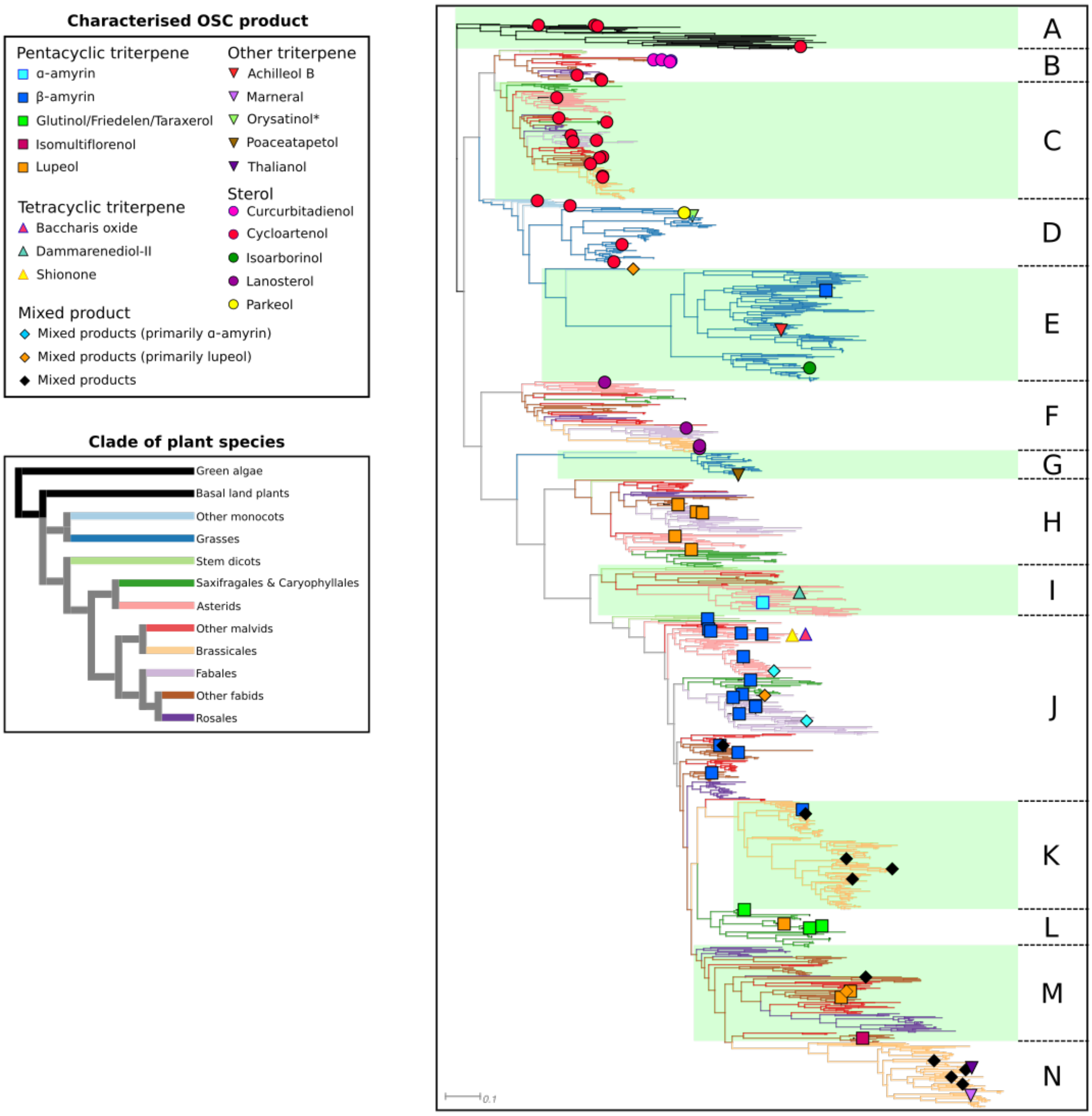


Figure 3.5 **OSC phylogeny from across the Viridiplantae**
Maximum-likelihood tree of 1404 OSC sequences mined from plant genomes and 82 characterised OSCs. Characterised OSCs are labelled according to the upper key and branch colours denote the plant clade to which the OSC sequence belongs (according to the lower key). OSCs are grouped according to letters (right), which often share functional specificity.

This phylogeny is consistent with previously published analyses [41,66], notably displaying the ancient gene duplication of the 'ancestral cycloartenol synthase (ACS) and the 'ancestral lanosterol synthase-like' (ALSL) [66], resulting in groups B-E and F-N, respectively. This will have occurred prior to the divergence of monocots and dicots approximately 140mya [66]. It is evident that the monocots have convergently evolved dammarenyl derived triterpene biosynthetic function via the ACS clade, versus the dicot OSCs which have achieved this via the ALSL clade.

Of the green algae and basal land plant species studied, all had a single OSC present in the genome, with the exception of *Selaginella* species ('spikemosses'). These are all represented in group A. Both *S. moellendorfi* and *S. kraussiana* appear to have OSCs present in tandem duplicates of either two or three at two distinct genome locations. The closely related Lycopodiaceae (clubmosses) are known to produce divergent triterpenoids via duplication and diversification of OSCs and squalene epoxidases [78], so it is likely that a similar range of OSC function would be found in *Selaginella* species.

Groups B and C represent all of the known dicot cycloartenol synthases (CASs). There is an apparent early duplication that precedes the divergence of the basal eudicots which results in the monophyletic B and C groups. Of the eudicots studied, 60% of species had one group B putative CAS, 30% had one group C putative CAS and 10% had both a group A and group B putative CAS. From a functional perspective, the key difference between these two groups is the presence of the cucurbitadienol synthase sub-clade in group C. These distinct OSCs have thus far only been characterised from cucurbits [16], although a range of plant species are presumed to produce cucurbitadienol given cucurbitacins and other cucurbitane-type triterpenoids are found across numerous monocots and dicots [79].

The monocot sterol and triterpene synthases are generally represented by groups D and E respectively, with the non-canonical orysatinol synthase [67] also present in group D and arborane-type sterol synthases in group E. There are two OSCs which fall between these two groups (and here are treated as basal to group E), one of which is a characterised mixed lupeol synthase from *Cheilocostus speciosus*.

The earliest OSC groups to diverge from the ALSL clade appear to be strongly conserved, being the dicot lanosterol synthases in group F and the monocot poaceaetapetol synthases in group G. The characterised OSCs within group H are all monofunctional lupeol synthases, however there appears to be numerous duplication and diversification events. This is typified by the presence of multiple representatives per genome, with a subset of these notable showing sequence divergence (indicated by increased branch lengths).

Furthermore, the apparent propensity for lupeol synthases to have diverged via multiple convergent evolution events across the dicots is noted, as well as their status as being OSC sequences basal to other triterpene biosynthetic OSCs in both monocots (group E) and dicots (group H). This could suggest that lupeol synthesis represents a relatively stable biochemical space for evolution to reach and/or that the selection pressures have periodically relaxed and increased for lupeol synthesis over evolutionary time across various clades.

Groups I-N represent what have historically been grouped together as diverse triterpene synthases [41,66]. Group I and groups J-N are two monophyletic groups, within which different plant taxons have representative sequences. Table 3.1 summarises the presence/absence of OSC groups across the dicot clades studied here.

Table 3.1 **Presence/absence of OSC groups across plant clades**
Green boxes represent the presence of an OSC group in the plant family indicated. Each land plant family shown here has a distinct set of such groups present in their genomes, demonstrating the diverse evolutionary paths OSCs have taken across the Viridiplantae. OSC groups are defined as in Figure 3.5.

Clade	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Green algae	🟩													
Basal angiosperms	🟩													
Monocots				🟩	🟩		🟩							
Stem eudicots		🟩	🟩					🟩	🟩	🟩				
Saxifragales/Caryophyllales			🟩			🟩		🟩		🟩		🟩		
Asterids			🟩			🟩		🟩	🟩	🟩				
Other malvids		🟩	🟩			🟩		🟩	🟩	🟩		🟩	🟩	
Brassicales			🟩			🟩		🟩			🟩			🟩
Other fabids		🟩	🟩			🟩		🟩	🟩	🟩			🟩	
Fabales			🟩			🟩		🟩		🟩				
Rosales		🟩	🟩			🟩		🟩		🟩			🟩	

The single unifying feature across these clades is that all of the genomes studied had at least one OSC present in either group J or group K (Brassicales only). These will be referred to as the 'core' triterpene groups. Furthermore, all of the characterised monofunctional beta-amyrin synthases (BAS) are present in these groups, and beta-amyrin is ubiquitously isolated from all plants [80]. These core groups are not monofunctional, but BAS sequences appear to be more conserved, with duplication and diversification appearing to drive alternate pathways, often via mixed-product synthases. Therefore, despite the variation in OSC function and diversity across the dicots, the evolutionary pressure to retain a functional BAS is evident from these data.

The dynamic nature of OSC diversification is evident, given the various duplication and loss events presumed to have occurred. For example, within the fabids, the Fabales generate all of their triterpene OSC diversity out of groups H and J, whereas the Rosales and other fabids also have a large number of representatives in the divergent group M (which also contains OSCs from asterids). The Fables have a larger repertoire of duplicated and diversified OSCs sister to BAS sequences within a single clade in group J. This may be expected, given this is their only known 'source' for non-lupeol triterpene synthases.

However, the effects of these evolutionary choices have on the subsequent triterpene 'biochemical space' these species have access to remains to be seen. Given the noted ability for OSCs to display convergent evolution (e.g. lupeol synthases) and reconstitution of diverse functionality via mutagenesis of small numbers of amino acids [67,69], it is possible be that plants are able to rapidly evolve any OSC functionality required regardless of their 'starting material'.

To summarise the above, Figure 3.6 is a cladogram of a proposed evolutionary pathway for the various OSC families mentioned here. The three earliest duplication events as described by Xue et al [66] are labelled.

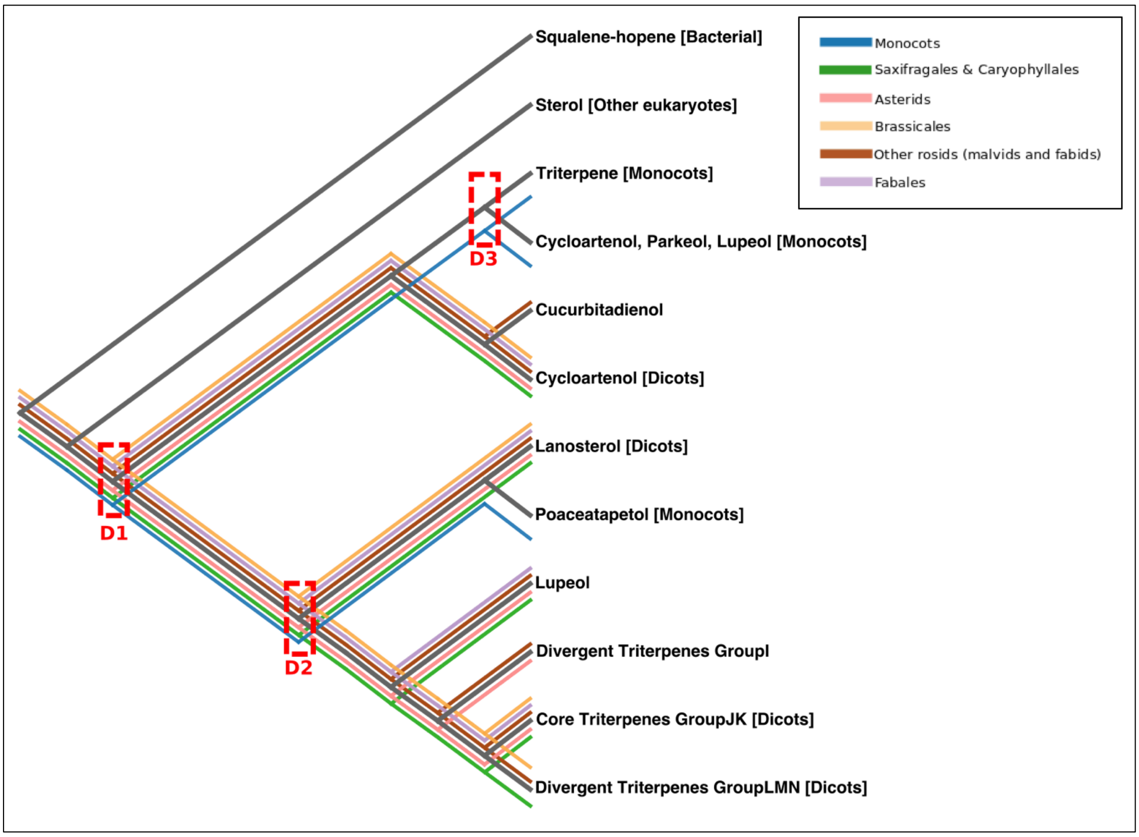

Squalene-hopene [Bacterial]
Sterol [Other eukaryotes]
Triterpene [Monocots]
Cycloartenol, Parkeol, Lupeol [Monocots]
Cucurbitadienol
Cycloartenol [Dicots]
Lanosterol [Dicots]
Poaceatapetol [Monocots]
Lupeol
Divergent Triterpenes GroupI
Core Triterpenes GroupJK [Dicots]
Divergent Triterpenes GroupLMN [Dicots]
D1
D2
D3
Monocots
Saxifragales & Caryophyllales
Asterids
Brassicales
Other rosids (malvids and fabids)
Fabales

Figure 3.6 **Cladogram of OSC evolution in plants**
Demonstration of the various evolutionary pathways OSCs have taken in different plant clades. Duplication 1 (D1) represents the ancient gene duplication of the 'ancestral cycloartenol synthase (ACS) and the 'ancestral lanosterol synthase-like' (ALSL) [66]. D2 and D3 as defined by [66] are also shown.

3.3.4 Profile-based classifications allow rapid screening of OSCs

The phylogenetic data generated here allow pHMMs to be generated for rapid classification of putative OSC sequences on-the-fly and therefore use in mining pipelines such as plantiSMASH and PhytoClust. Example data are shown in Figure 3.7, which also demonstrate the varieties in OSC complement across different plant clades as discussed above. The full tree is shown in Figure A1. This rapid annotation technique in the context with the functional and evolutionary relationship data discussed above is referred to as ‘OSC fingerprinting’, and it may be particularly useful in assessing candidates of interest for functional characterisation.

Given that groups J and M are paraphyletic, this approach is not perfect, and so care must be taken not to infer evolutionary relationships based on homology derived from these pHMMs without reference to a phylogeny. The impacts of poor genome assembly quality in the dataset can be observed given the infrequent occurrence of uncharacteristically low numbers of OSCs and/or high proportions of unclassified/pseudogenic sequences (e.g. *Psuedotsuga menziesii*, Figure 3.7A; *Lagenaria sicararia,* Figure 3.7F).

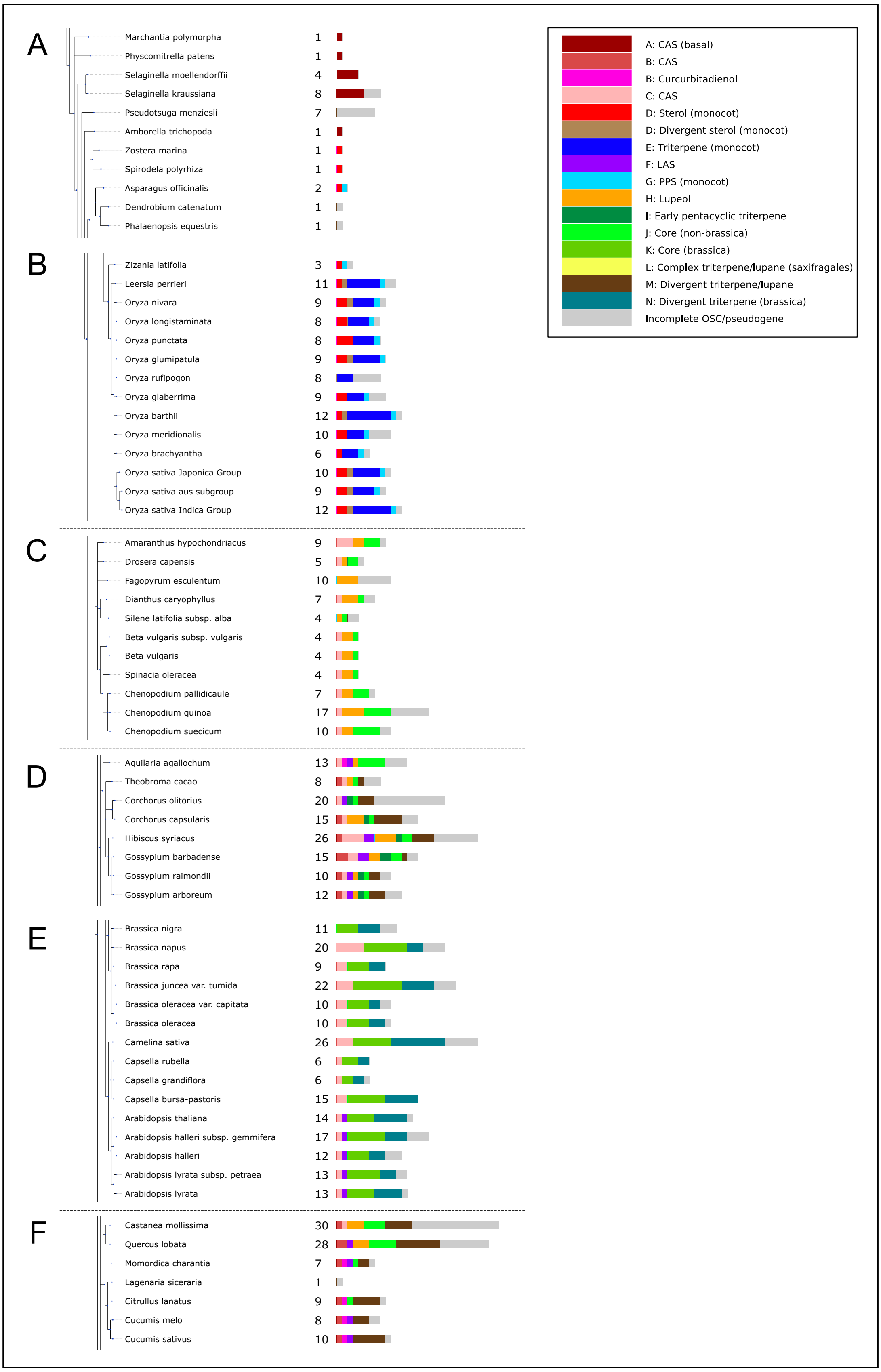


Figure 3.7 **OSC ‘genome fingerprinting’ across a variety of plant clades**.

Homology to conserved OSC groups can be used to predict the function of target candidates, discount candidates for desired functionalities and give snapshot as to the evolution and diversity of OSCs between species. A) Basal angiosperms and monocots B) Oryzeae C) Caryophyllales D) Malvales E) Brassicaceae F) Fagaceae and Curcurbitaceae. Full tree shown in Figure A1.

3.3.5 Functional predictions of uncharacterised OSCs

From this study, over a thousand uncharacterised, high-quality putative OSCs have been collated. It has been demonstrated that sequence-function relationships for the OSCs show varying complexity and that gene duplication and diversification appears to be a fundamental driving force for OSC evolution. These data therefore provide a clear opportunity for the selection, functional prediction and characterisation of OSC candidates.

Figure 3.8 shows three examples of candidates that were selected for characterisation. Gene sequences were verified against publicly available transcriptome data and synthesised by Integrated DNA Technologies (https://eu.idtdna.com/). Subsequent cloning, transient expression in *Nicotiana benthamiana* and product identification via gas chromatography with electron impact mass spectrometry fragmentation of leaf extracts was kindly carried out by Michael Stephenson (JIC). These examples demonstrate the ways in which phylogenetic relationships discussed above can be interrogated to choose candidate enzymes.

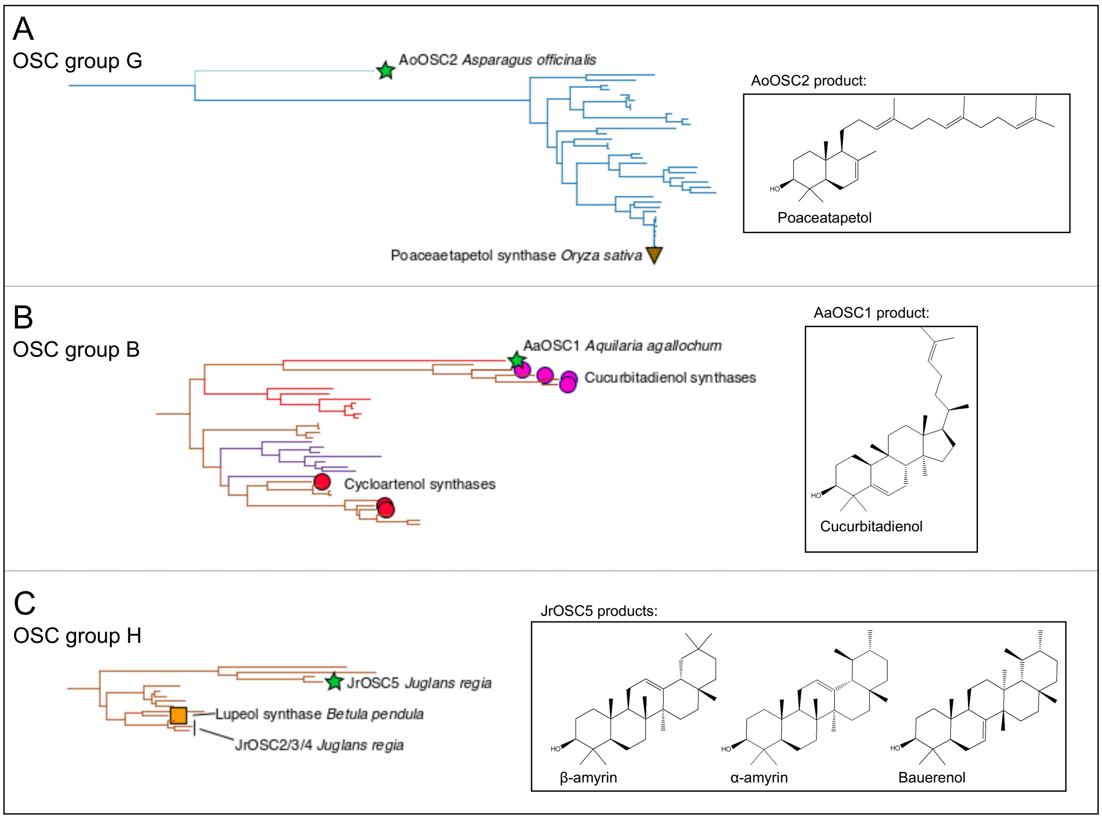


Figure 3.8 **Predicting OSC function by phylogeny**
Subtrees derived from Figure 3.5 to demonstrate sequences of interest and their function. Candidates selected for synthesis and functional characterisation signified by green stars. Functional characterisation kindly carried out by Michael Stephenson (JIC).

The poaceatapetol synthases are a recently characterised gene family [81] that appear to have strong functional conservation within the monocots, being ubiquitous across all species and with most having a single OSC homologous to this group (Figures 3.5, 3.7A, 3.7B, A1). Poaceatapetol is a pollen-specific triterpene which has been demonstrated to protect against dehumidification and was presumed to have evolved specifically in the Poaceae [81].

Figure 3.8A shows the poaceatapetol group G from Figure 3.5, including a putative gene from *Asparagus officinalis* (Asparagales), termed *AoOSC2*. Characterisation of the enzyme has identified it as a monofunctional synthase producing a bicyclic scaffold consistent with poaceatapetol or a closely related isomer (Michael Stephenson (JIC)). This demonstrates the conservation of OSC function within this group, shows that this function is not confined to the Poaceae and may indeed be ubiquitous to monocots.

Figure 3.8B shows a section of the group B OSCs (Figure 3.5) containing known cycloartenol synthases (CASs) and cucurbitadienol synthases (CCSs). Cucurbitadienol and derivative triterpene compounds are found across a wide range of plants, although are consistently produced by the cucurbits where they serve an anti-feedant role [16]. Furthermore, the only known CCSs have been found in cucurbits, however this study discovered a putative OSC in *Aquilaria agollochum* (Malvales) which shows homology (Figure 3.8B; *AaOSC1*).

A. agollochum, also known as agarwood, is a threated species which is known to produce a complex variety of terpenes including cucurbitacins I and E, is used in traditional Chinese medicine and is highly valued for its scented extract known as ‘oudh’ [82–85]. The draft genome sequence (which was analysed in this study) was previously analysed to find the candidate genes for cucurbitacin biosynthesis, but located only those for the upstream MVA pathway [85]. *AaOSC1* was discovered by the Selenoprofiles based approached described above. Characterisation of this enzyme has identified it as a CCS (Michael Stephenson (JIC)). As with *AoOSC2*, this is a demonstration of a functionally distinct OSC group. However, given that *A. agollochum* is in the order Malvales, the placement of these sequences in a monophyletic group could imply that this CCS family was present across all rosids and subsequently lost in the majority of species studied. Alternatively, it could be that it is due to both *Aquilaria* species and cucurbits both utilising the same ‘pool’ of evolutionary space in the group B CAS sequences to convergently evolve a CCS with shared sequence homology for chemical activity. Detailed analysis of these and related sequences is required to answer these questions further.

The genome of *Juglans regia* (common walnut; Fagales) was found to contain 13 OSCs, seven of which were assigned to group H (Figures 3.5, A1). Figure 3.8C shows part of group H containing a characterised lupeol synthase from *Betula nana* (Fagales) which three *J. regia* OSCs (*JrOSC2/3/4*) are sister to. Duplication and sequence diversification appears to have resulted in *JrOSC5*, which was characterised to encode a multifunctional pentacyclic triterpene synthase. Presuming the likely scenario that at least one of the JrOSC2/3/4 sequences is a monofunctional lupeol synthase, then this demonstrates not only duplication and sequence diversification, but a non-lupeol synthase in the group H OSCs.

3.3.6 Different OSC families show variation in propensity to be found in BGCs

The classification of OSCs across the Viridiplantae into groups based on the phylogenetic and functional data discussed provides an opportunity to revisit the propensity of different OSC families to fall into BGCs as classified by plantiSMASH 1.0 (Chapter 2). In the Brassicales, it is known that ‘Clade I’ OSCs (corresponding to group K) are not significantly clustered with CYP and acyl-transferase genes, whereas ‘Clade II’ OSCs (group N) are [35].

The relative frequency of the OSC groups across the whole genomes for the species analysed by plantiSMASH 1.0 (Chapter 2) were compared against the frequencies of occurrence within putative BGCs. These data are presented in Figure 3.9, where 3.9A shows the comparison of relative frequencies and 3.9B shows these data as a ratio.

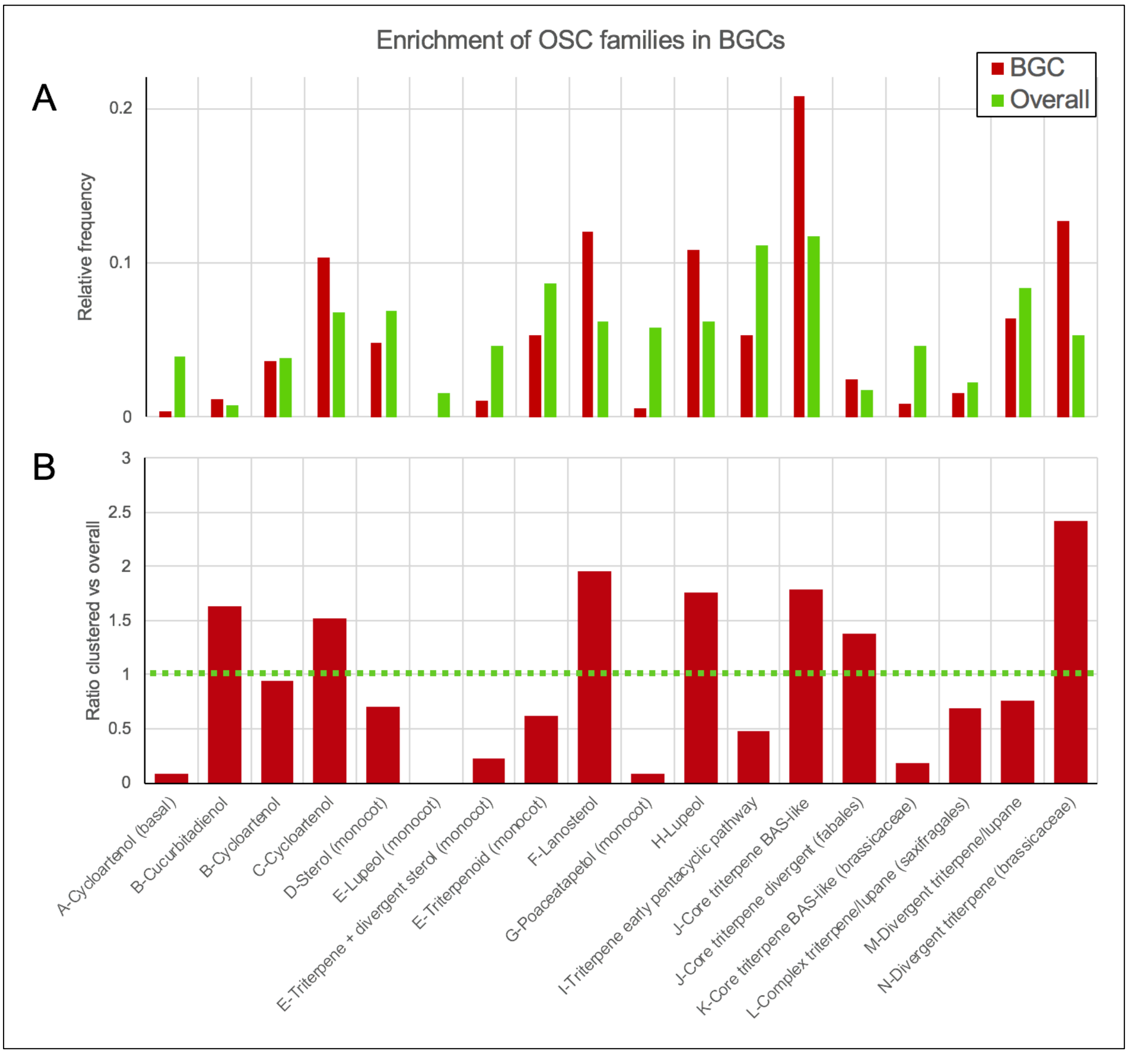


Figure 3.9 **Variation in propensity for different OSC groups to form BGCs**
A) Relative occurrence of various OSC groups within putative BGCs predicted by plantiSMASH 1.0 B) Ratio of BGC occurrence normalised to overall frequency of occurrence in plant genomes. Striking differences can be observed between the distribution of different OSC groups which are more or less often found in putative BGCs, suggesting some families have been ‘captured’ by the BGC formation process whilst others have not.

It is evident that there are striking differences in the propensity for these OSC groups to be present in putative BGCs. As expected from Liu et al [35], group N OSCs display a clear bias towards being clustered with other putative biosynthetic enzymes whereas the inverse is true for ‘core’ group K OSCs. Interestingly, the reverse case is presented for the non-Brassicaceae dicots where ‘core’ OSCs of group J are much more likely to found in BGCs compared to the divergent OSCs of groups L and M. This implies that the Brassicaceae have taken a particular evolutionary route for triterpene BGC formation and opens up broader questions with regards to what impact this has had on the prevalence of BGCs across the dicots.

3.4 Conclusions

The Selenoprofiles based mining approach developed here has been extremely successful, in not only approximately doubling the size of the OSC pool for phylogenetic analysis, but also in being accurate enough to allow functional characterisation of hitherto unknown protein sequences directly via gene synthesis and transient expression.

Numerous examples of apparent gene duplication and subsequent diversification are evident from the inspection of the phylogenetic data generated here and mixed product OSCs are often found as a result of this (Figure 3.5). It is possible that mixed product OSCs represent evolution ‘in progress’ where selection for a particular product has yet to optimise the pathway. Of course, it is entirely possible that a mixed product synthase would be selected for in specific circumstances. These questions will not be able to be answered until more is known about the various roles of triterpenes across plants. These can be highly complex and interconnected, as demonstrated by the recently discovered root microbiota modulation network in *A. thaliana* [1]. At a higher level, it is presumed that the formation and maintenance of triterpene BGCs will also interplay with OSC evolution, and the propensity for specific OSC families to be physically clustered offers a glimpse into the building blocks plants apparently utilise to achieve this.

As a model for exploring evolutionary sequence-structure-function relationships, OSCs are extremely well positioned. They are neither a rigid, conserved set of monophyletic and functionally distinct families, nor are they so dynamic that the link between phylogenetic clade and function is broadly lost, as is so often the case for other terpene synthase enzymes [6,27,28]. This may be linked with the chemical nature of their action, in that many of the cationic intermediates for triterpene biosynthesis act in a ‘cascade’ such that production of one compound over another can require the prevention of a specific reaction as much as the promotion of one [41]. This concept is sometimes termed ‘negative catalysis’ [86–88]. The nuanced nature of triterpene cyclisation is of course fundamental to the interest in its study, because nature has exquisitely solved reactions that are particularly challenging to access via conventional chemistry [21,26].

It has been shown here that functional predictions can be made via inference from phylogeny, but a clear opportunity exists to explore the sequence diversity presented here in greater depth, with specific focus on residues that are critical for determining OSC function. A preliminary unbiased investigation for specificity determining residues (SDRs) was carried out using the dataset generated here but these results were inconclusive (data not shown). Nonetheless, a number of residues been previously characterised as critical for functional specificity [41,67,69] and modelling approaches have shown some success in determining likely pathways for triterpenoid cyclisation for various OSC sequences [68,89]. Therefore, a more rational approach involving focussed studies of specific pathways and genes with rigorous analysis is likely to yield better results. The power of this approach will be increased with cross-reference to natural product database mining (e.g. Reaxys [90], Sci-Finder [91]) and integration of functional activities with protein modelling and structurally guided classification, as exemplified by CATH-DB [92].

Chapter 4. Cytochrome P450 classification and co-localisation with OSCs

4.1 Introduction

Cytochrome P450s (CYPs) are one of the largest enzyme families across all life and are critical in a broad range of metabolic and physiological systems [3,29]. They are a fundamentally important enzyme family in the majority of triterpene biosynthetic pathways in oxidising the scaffold at specific positions, functionalising it for biological activity and further modifications [33,41]. CYPs are classified by sequence homology into subfamilies, families and clans via alignment and expert inspection of phylogenies with already classified families. For example, CYP705A5 is in the "CYP705" family and is the 5th member within the "A" sub-family, whilst all CYP705 sequences are part of the "CYP71" clan (see [29] for a comprehensive review of plant CYPs). Nomenclature guidelines suggest a 40% sequence identity cut-off to classify a protein sequence into a given enzyme family and a 55% cut-off for subfamily. However if the protein sequence identity is below 60%, phylogenetic analyses are generally required [93].

The sequence-function relationship in CYPs is very dynamic; individual enzyme subfamilies display a wide range of bioactivities [29,94]. This means that there is little scope for predicting CYP function based on sequence data alone (as demonstrated with OSCs in Chapter 3). However, CYP genes are known to often be co-localised in plant genomes with OSC genes [36] and as such are almost ubiquitously found across the characterised plant BGCs so far [32,33].

Previous studies have reported not only significant OSC-CYP gene co-localisation across genomes of various plant species, but have also indicated that there may be a fundamental difference between monocot and dicot co-evolutionary dynamics [36]. Furthermore, recent analyses have demonstrated that the co-location of OSC and CYP genes in Brassicaceae species is highly dynamic, in that superficially homologous BGCs have in fact evolved independently [35]. OSC and CYP genes are therefore frequently co-functional and the corresponding genes are often co-localised in the genome. There is a natural opportunity to use the genome data collated here to broaden studies of the genome organisation of OSC-CYP gene pairs. However, to achieve this, a high-throughput and systematic annotation method will need to be implemented as plant genomes typically contain hundreds of CYP genes [29]. The number of CYP genes requiring classification across large-scale genome mining would be therefore be unmanageable to perform manually.

4.1.1 Aims:

The aims of this chapter are to:

- Develop an accurate and rapid classification methodology for CYP protein sequences

- Perform OSC-CYP gene co-localisation enrichment analyses across a wide range of suitable plant genomes
- Investigate these results to observe what, if any, patterns are found across a wide range of plant species and what implications this has for OSC-CYP gene co-evolution, plant BGC formation and functional predictions of CYP sequences

4.2 Methods

4.2.1 Database generation

Sequences and annotations of 11215 CYPs of 61 plant species were manually downloaded and organised from the curated P450 homepage [93] as high-quality representative sequences for plant CYPs. These sequences were added to a database of over 51000 CYP sequences automatically downloaded from CYPED [95], which contains semi-automatic classifications of sequences across prokaryotes and eukaryotes. After filtering for duplicates and low-quality sequences a database of 60056 CYP sequences was generated, all of which with putative family annotations and formatted for downstream use with CD-HIT [96].

4.2.2 CD-HIT-based scoring

To develop a suitable clustering threshold with CD-HIT, a jackknife resampling approach was taken. Specifically, the total sequence database was randomly split into 90% training data and 10% test data. The training data was clustered with CD-HIT at a given sequence identity threshold, resulting in a set of representative sequences for each cluster. For each cluster, the proportion of CYP families that made up each cluster was calculated. The test data was then added to the representative sequences and CD-HIT was run again at the same threshold used previously. For each CYP sequence in the test data, the cluster that it was assigned was reported thus determining the putative family classification. This process is summarised in Figure 4.1.

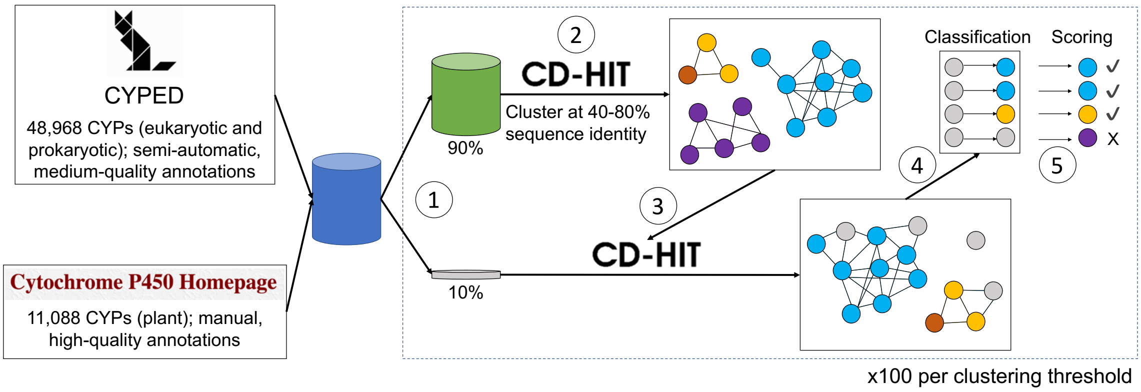

CYPED
48,968 CYPs (eukaryotic and prokaryotic); semi-automatic, medium-quality annotations
Cytochrome P450 Homepage
11,088 CYPs (plant); manual, high-quality annotations
1
2
CD-HIT
Cluster at 40-80% sequence identity
90%
10%
3
CD-HIT
4
Classification
Scoring
5
✓
✓
✓
X
x100 per clustering threshold

Figure 4.1 **Computational workflow for CD-HIT clustering of CYP database sequences**
High quality plant CYP sequences from P450 homepage [93] and lower quality sequences from CYPED [95] were collated and manually organised to generate a core database of protein sequences classified into families. Various sequence identity thresholds were tested to optimise the CD-HIT based CYP classification tool, with 100 iterations per threshold in a jackknife resampling approach using 90% as training data and 10% as test data. Sequential steps are numbered.

To demonstrate; if a test CYP sequence fell into a cluster made up of entirely of a single CYP family, the annotation would return confidently as that family. If a sequence fell into a cluster comprising a mix of families, then the annotation would return as a proportional score representing the families in the cluster. If a CYP fell into a cluster made entirely of other test sequences, the annotation would return as unknown. Scoring was then carried out depending on how many CYPs were annotated correctly, with ambiguous results returning the family with the highest proportional score. This was carried out 100 times and the results used to set the clustering threshold, as well as indicate the efficacy of this annotation approach.

4.2.3 OSC-CYP gene enrichment

To ensure results were not biased due to low quality genome assemblies, a subset of 88 high-quality genomes from 60 plant species were used for the OSC-CYP co-localisation analysis. OSCs were located via HMMer using a profile comprised of 82 characterised OSC sequences [41]. The neighbouring genes were identified in an envelope of +/- 10 genes upstream and downstream of the OSC. A gene count was used because of the variable gene density found in plant genomes. The Pfam profile PF00067 and the CD-HIT CYP-classifier was used to define and subsequently classify the CYPs found into families. Enrichment analysis was performed using the Fisher’s exact test between the proportion of CYPs of each family and clade in the neighbourhood envelope and the corresponding counts across the entire genome. Figure 4.3 displays a summary of this approach.

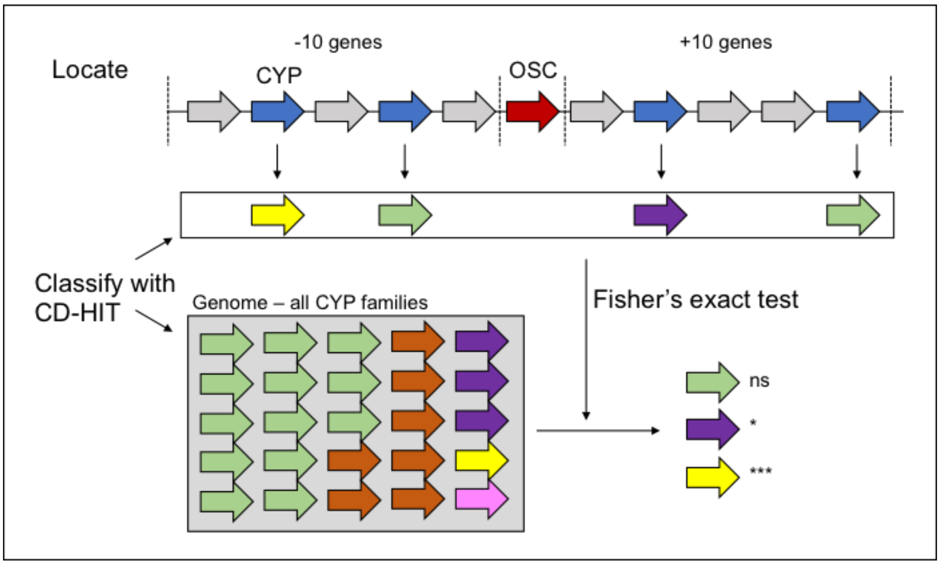


Figure 4.2 **Summary of OSC-CYP co-location methodology for neighbourhood enrichment** OSCs and CYPs were located in each target genome using HMMer. CYPs were then classified according to the methodology described above. The frequency of occurrence of each CYP family within 10 genes of the OSCs was compared to the overall frequency in the genome and Fisher's exact test used to assess if any given CYP family was significantly enriched in these areas.

4.3 Results

4.3.1 CD-HIT clustering is fast and accurate

Approximately 60,000 CYP sequences were collated for use in homology finding and family annotation, comprising high-quality plant CYP data from the "Cytochrome P450 Homepage" [93] and lower quality data from CYPED [95]. Initially a profile-based method was developed, involving the creation of specific pHMMs for each CYP family and alignment scores used to determine annotation. However, this approach was resource intensive and very sensitive to any incorrect annotations present in the original database, producing an unacceptably variable output (data not shown).

The sequence clustering tool CD-HIT [96] was chosen instead to classify CYPs due to its speed in clustering gene sequence data based on sequence identity, and because sequence identity is a key metric by which CYPs are manually classified. Furthermore, CD-HIT is part of plantiSMASH 1.0 [34], therefore its use would simplify any potential future integration.

Figure 4.1 demonstrates the methodology used for testing and optimising a classification approach using CD-HIT. To optimise CD-HIT for these data, test data was generated by randomly sampling 10% of the sequences from the collated database. These were then clustered with the remaining annotated data and assigned to families based on the majority of sequences present in the clusters each sequence fell into. This was repeated 100 times for a range of sequence identity thresholds for clustering. Two key benefits of this method over a profile-based, alignment approach is that edge-cases are able to be returned as unclassified and a small number of misannotations do not have a large 'knock-on' effect. The summary statistics of this jackknife resampling approach are shown in Figure 4.3.

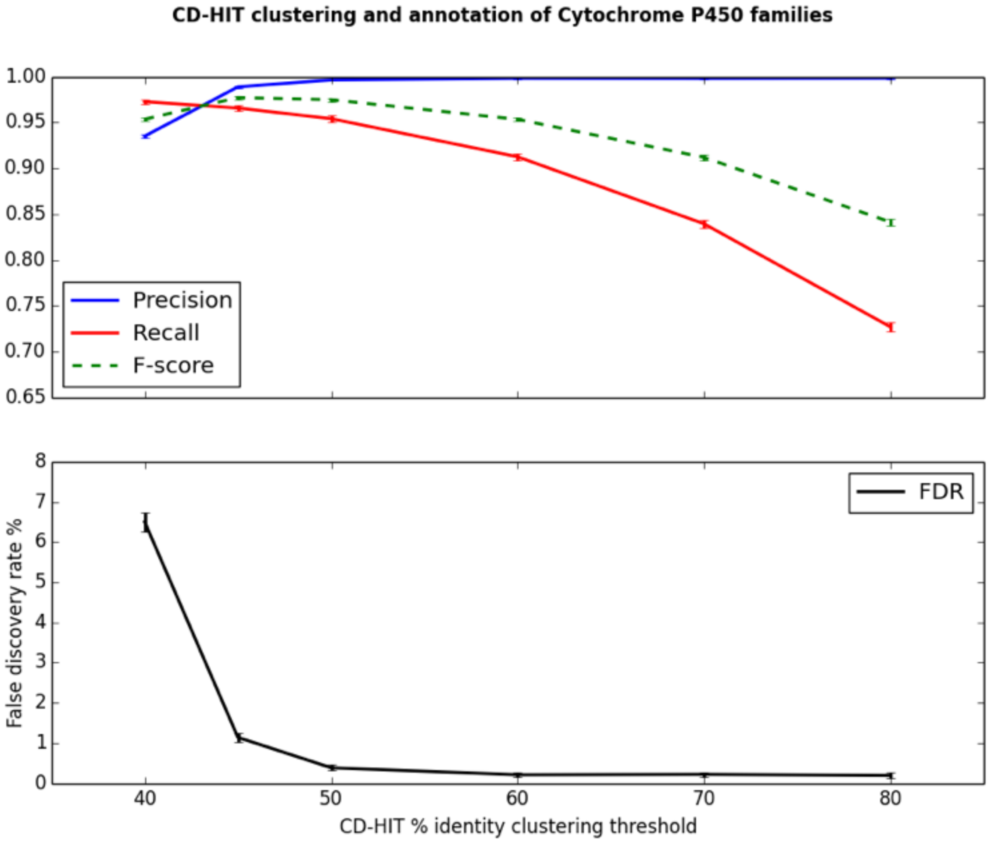


Figure 4.3 **Summary statistics for results of testing CD-HIT CYP classification approach** (Top) Precision, recall and F-statistic. (Bottom) False discovery rate. A low clustering threshold results in loss of precision, where CYPs from different families are grouped together, as demonstrated by the sharp increase in false discovery rate below 45%. As the sequence identity threshold is increased, only those sequences with high homology to the training data are classified, demonstrated by the decline of the recall statistic (i.e. increasing false negatives).

As can be seen from Figure 4.3, clustering at 60% sequence identity and above returns a near-perfect precision in assigning CYP genes to families. This is in-line with expectations, given below this value phylogenetic data are generally used to verify family membership. Low recall values at high thresholds is also expected, as families are broken into sub-clusters which are disconnected from annotated sequences. For implementation of this tool, the 50% threshold was decided to be optimal. Whilst the 45% threshold returned a marginally higher F-score, an increase in the false discovery rate (incorrect classifications) is less preferable than a slight increase in false negatives (unclassified sequences) for the purposes of this study.

To demonstrate the power of this approach, the 120 CYPs discovered to be co-located with terpene synthases in a previous study [36] were chosen. These sequences were removed from the starting database and then assigned at a 50% sequence identity threshold. This resulted in 91% of sequences being correctly annotated to a family level, 5% to a clan level, 4% were returned as uncertain and <1% were incorrect. Because this approach allows thousands of sequences to be accurately assigned to families in seconds, it immediately provides an opportunity to carry out high-throughput co-evolutionary studies for OSCs and CYPs across the plant kingdom.

4.3.2 OSC neighbourhood enrichment shows clade-specific conservation of CYP families co-localised with OSCs

Figure 4.2 summarises the approach used for testing enrichment. The genomes of 60 plant species were used for this study (based on suitable assembly and annotation quality, Table A1). For each OSC gene located via HMMer, the ten genes upstream and downstream were sampled and any CYPs classified according the method described above. The frequency of families co-located with OSCs were then compared to the overall distribution of those families across the whole genome. These counts were compared using Fisher's exact test and the resulting *p*-values used to determine significance of CYP family enrichment. *p*-values below 0.05 were counted as significant, though given the strictness of Fisher's exact test and the potentially very low counts of CYP family members across a genome, a record was also made in cases where $p<0.1$ for further investigation.

Figure 4.4 shows a full taxonomy of the species analysed and the most frequently occurring CYP families and clans found to be co-located with OSCs. The majority of Angiosperms show significant OSC-CYP co-localisation, however not all do. *Malus* species, *Prunus* species, *Capsicum* species and *Arachis* species all have no significant OSC-CYP pairing, though the sister species within their clades do (Figure 4.4). Putative BGCs are returned by plantiSMASH 1.0 for these species and many of these include CYPs, but very few include OSCs. This indicates that gene clustering may still be a mechanism utilised by these species, but not for early triterpene biosynthesis. *Cuscuta australis* similarly shows no OSC-CYP co-localisation (Figure 4.4), though this species has undergone significant gene loss and genome reduction due to its parasitic lifestyle (as well as loss of roots and leaves) [97].

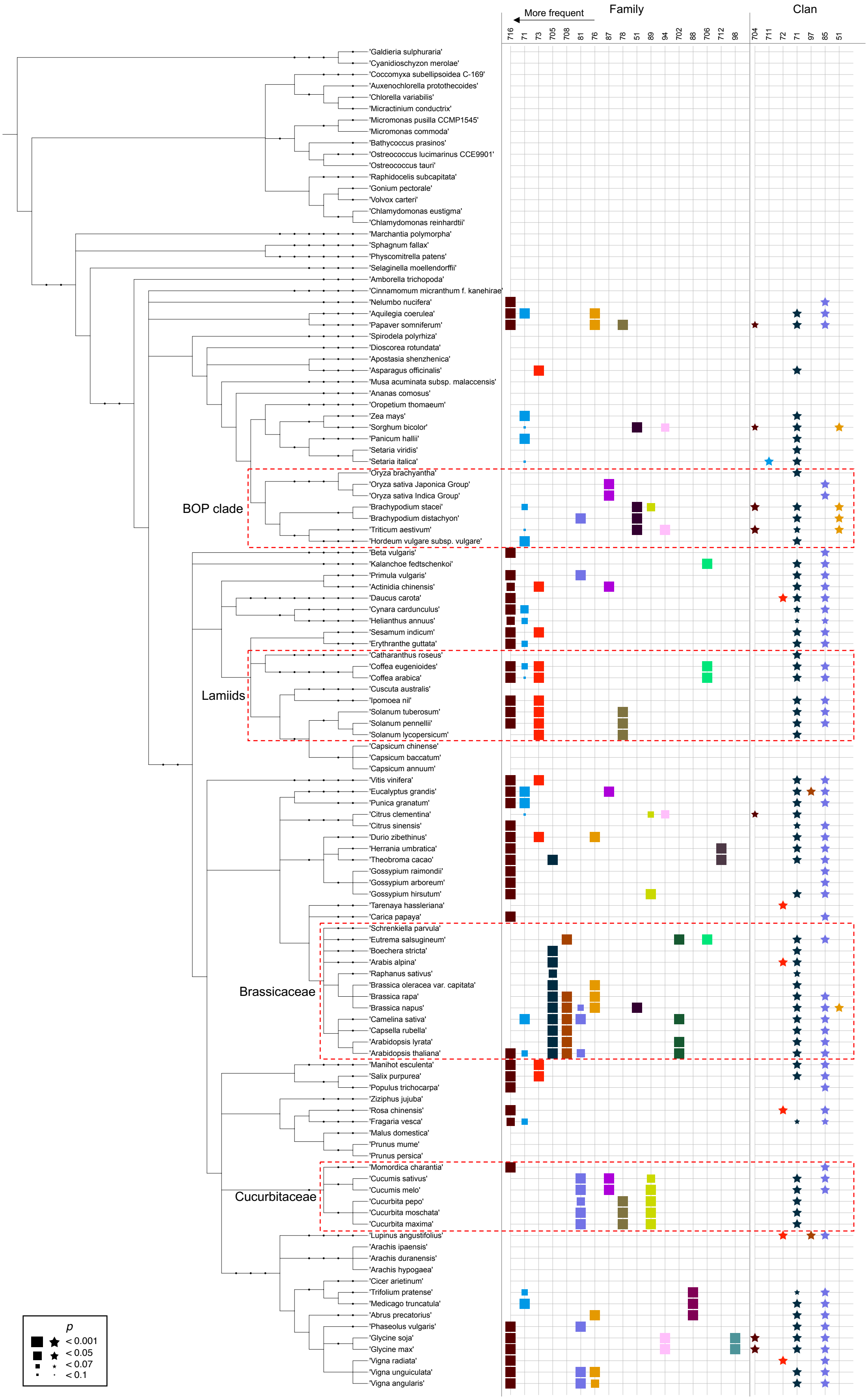


Figure 4.4 **CYP families significantly enriched in OSC gene neighbourhoods**

Shape size represents significance determined by Fisher's exact test, shown in key. CYP families/clans are ordered by decreasing frequency of occurrence (left to right). Clades discussed in the text are highlighted by red boxes.

Across all the species analysed, the CYP716 family was most frequently found to be significantly co-located with OSC genes. This is encouraging, given the numerous examples of CYP716 enzymes showing functionality in triterpene biosynthetic pathways [1,33,41]. Furthermore, inspection of specific plant clades in Figure 4.4 demonstrates there is a notable level of conservation of CYP gene families found co-localised with OSC genes. For example, the CYP51 family in *Brachypodium* and *Triticum* species, the CYP73 family amongst the lamiids, the CYP705 and CYP708 families in the Brassicaceae, and the CYP81 and CYP89 families in the Cucurbitaceae. Many of these families are known to functional members of triterpene BGCs in the relevant species [1,16,33] (Figure 4.5), which gives good support for this approach highlighting not only co-localisation patterns, but functionally relevant relationships.

Previous studies have postulated a fundamental difference between monocot and dicot CYP co-localisation patterns with terpene synthase genes, wherein dicots show conservation in clan type (primarily CYP71) in terpene synthase gene-*CYP* pairs and monocots have a wider range of CYP clans associated across all terpene synthase families [36]. Figure 4.4 demonstrates that, given the numbers of species analysed, the monocots do have a proportionally more diverse range of CYP clans co-localised with OSCs, but almost all CYP clans are represented in both the monocot and dicot data. These data generally indicate that CYP gene family recruitment is dynamic across angiosperms, for example, *Glycine* species are unique amongst the fabids in recruiting CYP704 clan genes.

Whilst this analysis did not encompass all terpene synthase families, it is argued that previous conclusions were biased due to low sample numbers of available genome data, which this study has been able to overcome. These data overall demonstrate the ability for plant species to dynamically recruit different CYP families regardless of clade, but that conservation of *OSC-CYP* co-localisation within a plant clade exists and may reflect conserved functional activity. Practically, this clade-specific conservation may aid in the selection of candidates for functional characterisation, in that certain CYP families may be targeted for likely activity based on the co-localisation patterns seen in a given species relatives.

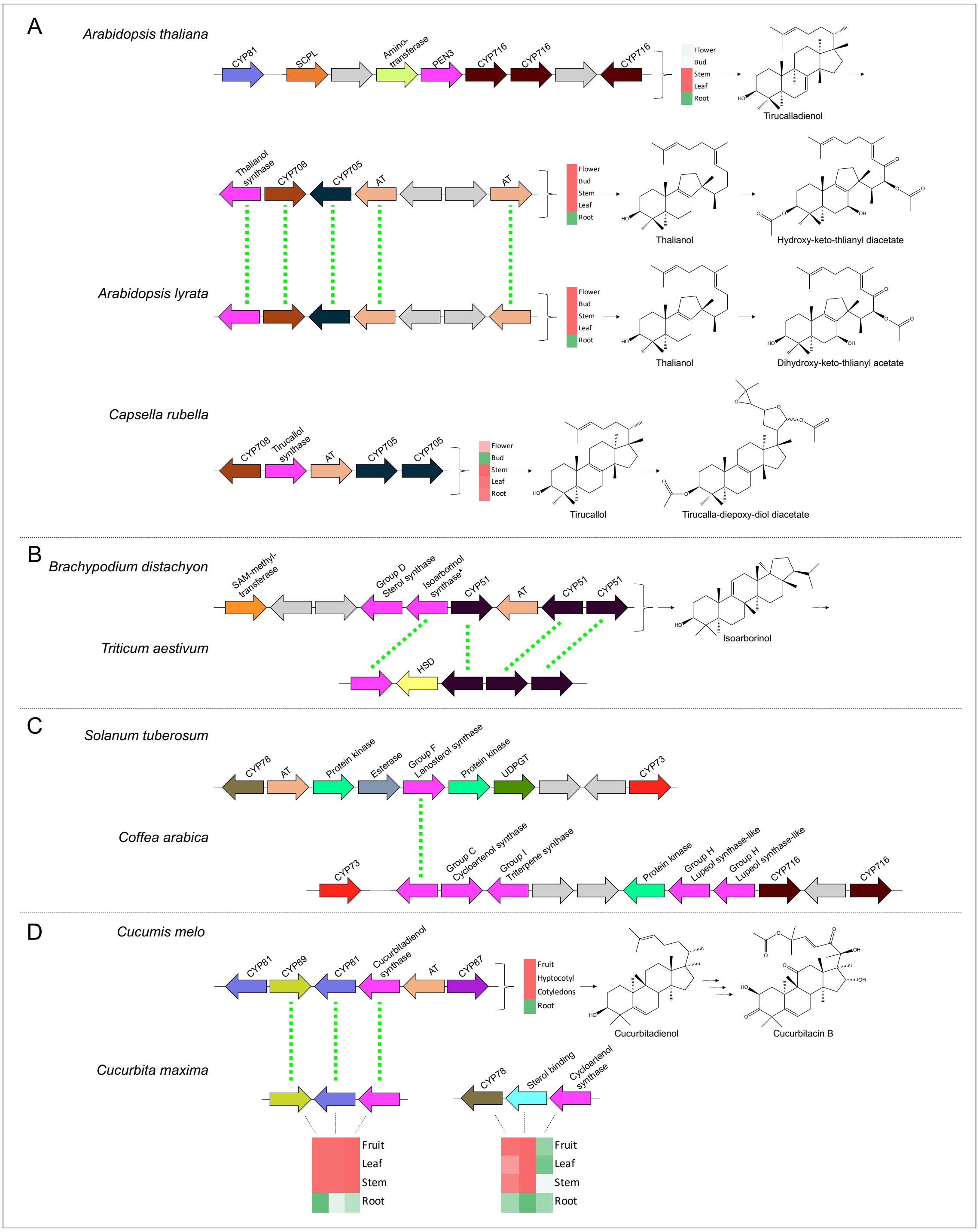


Figure 4.5 **OSC-CYP co-location captures various gene environments**
Examples of gene regions around OSCs where CYP families were found to be significantly enriched. Green dashed lines represent orthology between species. Expression data derived from [1] (A) and [85] (D).

4.3.3 Closer study reveals variation within clades demonstrating CYP family conservation

Despite showing superficial orthology (based on the conservation of CYP families), recent analyses have shown that some BGCs within the Brassicaceae have originated independently. Specifically, Liu et al. showed with careful phylogenetic inspection and characterisation of the *Arabidopsis sp*. thalianol BGC and the *Capsella rubella* tirucallol BGC (Figure 4.5A) that, while each BGC is functional and contains a group N OSC, a CYP705, a CYP708 and a BAHD acyl-transferase, they do not share an 'ancestral BGC' [35]. Instead, it is postulated that these BGCs have formed independently via the recruitment of genes derived from the shared Camelineae ancestral karyotype.

Figure 4.5A summarises these data, where orthology is shown with green dashed lines. This shows that the *Arabidopsis* species produce thalianol-derived triterpenes using CYP705 and CYP708 family enzymes, whereas the *C. rubella* BGC encodes a tirucallane-derived pathway. The data presented here also demonstrate that *A. thaliana* is unique within the Brassicaceae in recruiting the CYP716 family to *OSC* loci (Figure 4.5C). This is notable, given that across the angiosperms, CYP716 was the most frequently co-located CYP family but is otherwise absent within the Brassicaceae. Furthermore, the BGC containing CYP716 family genes encodes a pathway beginning with the production of tirucalladienol (Figure 4.5A). It is not known whether this pathway shares functional homology to that found in *C. rubella*.

Overall, the deceptively intricate relationship between the BGCs of *Arabidopsis* species and *C. rubella* demonstrates how care must be taken in assuming orthology based on the conservation of gene families, especially given the implications that such assumptions may have for downstream evolutionary and functional analyses.

To demonstrate further variation in putatively conserved OSC-CYP co-localisations within plant clades, examples are given in Figures 4.5B-D. The conservation of CYP51 family recruitment to OSCs in *Brachypodium* species and *Triticum aestivum* (Figure 4.4A is due to orthology (Guy Polturak (JIC), Figure 4.5B). Furthermore, the *Brachypodium* species BGC is known to produce isoarborane derived triterpene compounds from a group D isoarborinol synthase (Guy Polturak (JIC), Figure 4.5B).

CYP51 family enzymes are also present in the *Avena strigosa* avenacin BGC (specifically the *CYP51H10/Sad2* gene). *A. strigosa* is also a member of the Pooideae. However, this BGC is not homologous to those shown in Figure 4.5B, with the group E beta-amyrin synthase catalysing the first step of the avenacin biosynthetic pathway. Elsewhere in the monocots, *Sorghum bicolor* is the only other species that demonstrates enrichment of CYP51 family enzymes co-located with OSCs (Figure 4.4). *S. bicolor* is relatively taxonomically distant in relation to *Avena* and *Triticum*. These data therefore imply a minimum of two, and possibly three, separate instances of monocot species evolving and maintaining functional OSC-CYP51 pairings, along with further BGC expansion.

The lamiids show conservation of CYP73 and CYP716 family co-location (Figure 4.4) but the examples from *Coffea arabica* and *Solanum tuberosom* in Figure 4.5C demonstrate the variability in the OSC genomic neighbourhood. In both of these regions, a group F lanosterol synthase is present. However, *C. arabica* also has four further OSCs and a total of four divergent OSC groups at the same locus (Figure 4.5C). These are not tandem duplicates (barring the two group H OSCs), as they are from functionally and evolutionary separate clades (Chapter 3).

Furthermore, given that the group F lanosterol group is a monophyletic clade, and both *C. arabica* and *S. tuberosum* only have one representative from this clade within their respective genomes, it is reasonable to consider these genes to be orthologous.

The genes at these loci have not been characterised and the role of lanosterol in plant metabolism is poorly understood [41,98]. Therefore, co-expression data would be especially useful in further analyses of these genomic regions. Nonetheless, the results reported here demonstrate how variable such *OSC-CYP* gene co-location can be, even within a plant clade that has conserved CYP family recruitment (Figure 4.4).

The cucurbitane triterpenes found in the cucurbits are diverse and previous studies have demonstrated the convergence and divergence of constitutive BGCs, biosynthetic enzymes and regulatory genes in *Cucumis* species and *Citrullus lanatus* [16,99]. Both *Cucumis* species and *Cucurbita* species show conservation of CYP81 and CYP89 family OSC-CYP gene co-location, yet differ in their utilisation of CYP87 (specific to *Cucumis* species) and CYP78 (specific to *Cucurbita* species) family genes (Figure 4.4).

Figure 4.5D illustrates the characterised BGC in *C. melo* which is known to encode the initial steps of cucurbitacin B synthesis [16]. *Cucurbita* species are also known to produce a range of various cucurbitanes [16,79], and *C. maxima* contains an orthologous BGC containing a CYP89 gene and a CYP81 gene co-located with a group B cucurbitadienol synthase. As expected, given Figure 4.4, the *C. melo* BGC also contains a CYP87 gene which is not present in *C. maxima*. However, the position of a CYP78 gene in *C. maxima* is not within the putative cucurbitane BGC, but instead is co-located with a putative sterol-binding family gene (PF00173) and the *C. maxima* characterised cycloartenol synthase (Figure 4.5D). Inspection of co-expression data shows that whilst the cycloartenol synthase is expressed across the tissues of the plant, the sterol-binding and CYP78 genes are root-specific, as is the cucurbitane BGC. *C. maxima* is known to produce various cucurbitacins (including cucurbitacin B), and the full cucurbitacin pathways within the Cucurbitaceae are not encoded on a single BGC [16]. It is therefore possible that this CYP78 and sterol-binding gene are functionally relevant to the cucurbitacin biosynthetic pathways in *C. maxima*.

These examples are presented here to demonstrate the variability in *OSC-CYP* co-location patterns within plant clades that appear to be conserved at a CYP family level. These data are proof of the ability of plant species to retain a ‘pool’ of CYP enzymes that can be utilised for specialised metabolism, but that this relationship is very dynamic and can undergo rapid diversification.

4.4 Conclusions

This chapter demonstrates the necessity and impact of rapid and accurate classification tools in high-throughput mining pipelines. The clustering-based approach developed here allows broad systematic analysis of CYP families where previously manual inspection and annotation would have been required. This methodology has produced results with consequences for both the evolutionary dynamics of OSC-CYP co-localisation as well as providing avenues for candidate selection and functional predictions in enzyme characterisation.

The consistency with which CYP genes are significantly co-located with OSCs across all angiosperm taxa is of particular importance, given how relatively little is known of the broad scope of plant BGC formation and diversity. This approach is not limited to the definitions of BGCs as discussed in Chapter 2, and so provides a more granular study into plant genomic organisation between two gene families which are, importantly, consistently co-functional. Furthermore, the propensity of different plant clades to utilise a characteristic 'pool' of CYP families is notably similar to the patterns observed in OSC family distribution in Chapter 3. Given that these co-location data correlate with already characterised BGCs and OSC-CYP pairs [36], these data generally provide a compelling case for the widespread utilisation of gene co-location for functional co-regulation in specialised metabolism across all angiosperm species.

Nonetheless, it has also been observed the evolutionary tool of gene clustering is highly dynamic in plants, with closely related species independently assembling OSC-CYP pairs despite drawing from same 'pool' of OSC and CYP families [35] (Chapter 3). It has also been noted that certain species do not appear to have any OSC-CYP co-localisation at all. The ability for individual species to recruit, maintain and remove co-localised OSC-CYP genes therefore appears to be both ubiquitous and relatively rapid, and there is no evidence to suggest this process is limited to the gene families studied here. As more is discovered with regards to the functional role and regulatory networks of these specialised metabolites [57], a more comprehensive framework can be generated to explain what evolutionary benefit such organisation provides.

In terms of predicting functionality, classification of CYPs into families determined by sequence homology does not offer the same opportunities for direct, sequence-based predictions as with OSCs. Across triterpene biosynthetic pathways, CYP functionality is highly diverse [29,94] and in wider metabolic research, intense modelling and machine learning approaches have so far shown limited success in predicting CYP function from sequence and/or structure in very specific cases [100–102]. Nonetheless, it has been previously noted that CYP716 enzymes are particularly rich in triterpene activity [103], which this study has supported given that CYP716 family genes were the most frequently co-located with OSC genes across all plants.

Furthermore, specific CYP families have been shown to play key roles in the triterpene complements of certain species such as CYP705 and CYP708 in the Brassicaceae [1], CYP81 and CYP89 in the Cucurbitaceae [16], CYP72 in the legumes [104] and CYP51 in the monocots

[42]. This chapter has shown that indeed such patterns (and more) can be found across numerous plant clades. Therefore, by comparison with related species, some predictions can be made as to whether a given CYP is likely to be co-located and/or co-functional with an OSC using only transcript sequence data. Given that plant genomes generally contain hundreds of CYP genes, this may be useful in narrowing the search space. Other data, such as gene expression or metabolite analysis, can be combined with these for further refinement.

The power of the computational approach developed here will only grow as it is applied to larger sets of genome data, but opportunities exist already to pick apart the OSC-CYP relationship further. Firstly, a CYP subfamily classification approach is likely to be possible, which may provide more evidence as to which CYP sequences are subject to selection for recruitment. Model refinement would also build scope for highlighting potentially novel CYP subfamilies without intensive phylogenetic study. Secondly, this approach may be combined with the OSC profiles generated in Chapter 3 to determine if there is a more nuanced relationship at play. Evidence for this is already strong, given the varying frequency with which different OSC subfamilies are found in putative BGCs (Figure 3.6) and that only group K OSCs were found to be significantly clustered with CYP705 and CYP708 sequences in certain Brassicaceae species [35]. This neighbourhood enrichment approach may also be 'inverted' and applied to CYPs. Finally, outside of BGC dynamics, this CYP classification approach could be applied to produce an automatic and ongoing summary of CYP diversification and evolution amongst all plants, such as demonstrated by manual classification [29,93].

Chapter 5. Predicting GT1 substrate specificity in *Quillaja saponaria*

5.1 Introduction

5.1.1 GT1s

Glycosylated triterpenoids, also known as saponins, are of particular interest in metabolic study because of their propensity to be biologically active. This is due to their amphiphilic nature, allowing them to interact with biological membranes and act as particularly good surfactants [5,41,105]. Saponins have therefore found use in a variety of contexts, including as soaps, cosmetics, foaming agents, vaccine adjuvants, marine toxins and anti-feedants [5,12,14,51].

A key enzyme family responsible for the biosynthesises of saponins are the family 1 glycosyltransferases (GT1s), alternately called UDP-dependent glycosyl-transferases (UGTs), which have been characterised in numerous triterpene biosynthetic pathways [33]. GT1s utilise a UDP sugar donor and transfer the sugar moiety onto an acceptor molecule via an S_N2-like reaction mechanism [30]. The most common function of GT1s is as a glucosyltransferase, though GT1s with various sugar specificities have been discovered [30,41].

A recent review of characterised GT1s in plants has highlighted numerous aspects of their sequence-structure-function relationship [30]. A great deal remains unknown about how GT1s control substrate specificity, but certain trends have been identified. Firstly, sequence analysis of characterised plant GT1s shows that the acceptor molecule (e.g. flavonoid, triterpenoid) and/or the reaction function (e.g. ester forming, glycosidic branch elongating) can often be revealed by phylogenetics (Figure 5.1), though a detailed understanding of what residues are required for acceptor specificity remains unknown.

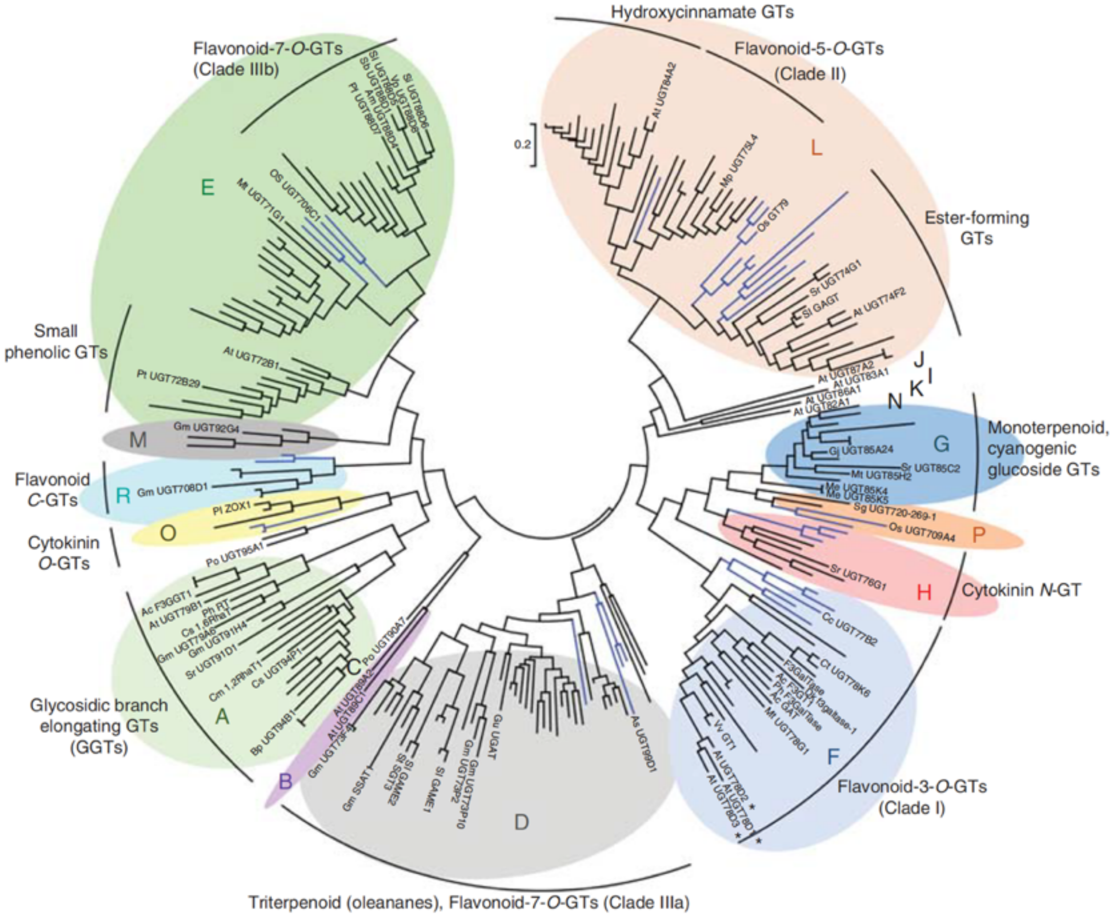


Figure 5.1 **Phylogenetic tree of characterized plant glycosyltransferases 1 (GT1s)** Reconstruction of GT1 phylogeny from a collection of 246 biochemically characterized GT1 protein sequences. The groups are delineated as defined by [106] and [107]. Figure and legend adapted from [30].

A number of residues have been determined to be relevant in the sugar donor specificity of GT1s [30]. A defining feature of GT1s is the presence of the 'plant secondary product glycosyltransferase' (PSPG) motif, a conserved sequence of 44 amino acids (Figure 5.2A) which is prominent in the sugar donor binding site of GT1s (Figure 5.2B) and variation in which has been shown to impart a degree of sugar specificity. For example, mutagenesis of the final residue from Q44 to H44 has been shown to confer specificity for galactose and arabinose over glucose [30]. It has also been found that GT1s that selectively utilise glucuronic acid have an R22 in the PSPG motif instead of the more usual W22 [30]. Another structurally conserved feature of GT1s relevant for sugar donor specificity is the N5 loop, mutagenesis of which has indicated that it confers selectivity between hexose and pentose sugars [30].

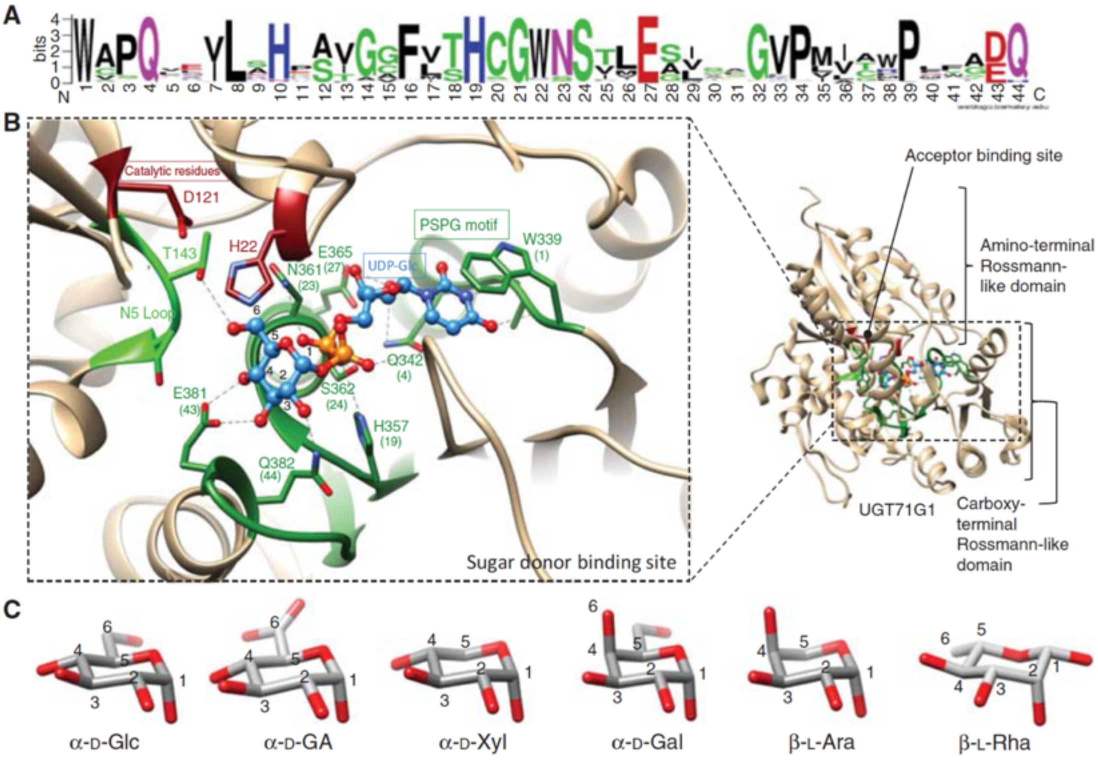


Figure 5.2 **Determinants of sugar specificity of plant glycosyltransferases 1 (GT1s)**
(A) Consensus plant secondary product glycosyltransferase (PSPG) motif generated by weblogo (weblogo.berkeley.edu) from an alignment of characterized GT1s. (B) The sugar donor-binding site for the crystal structure of the GT1 enzyme UGT71G (a flavonoid/triterpenoid O-glucosyltransferase Medicago truncatula) in complex with uridine diphosphate glucose (UDP-Glc) (PDB code 2ACW). The PSPG motif is shown in dark green, the N5 loop in light green, and catalytic residues in dark red. UDP-Glc is shown as a ball and stick model and colored in blue. Proposed hydrogen bonds are shown as dashed lines. (C) Compared structures of most common sugar donors of plant GT1s. Figure and legend adapted from [30].

5.1.2 QS-21

The soapbark tree, *Quillaja saponaria*, has been utilised since its discovery for its high concentration of saponins, traditionally as a detergent, a foaming agent and an expectorant, but also in modern medicine as an immunological adjuvant [12,13,108–110]. Saponins are often able to illicit an immune response, though many are unacceptably cytotoxic for clinical use. However, a specific saponin from *Q. saponaria*, termed QS-21, is used safely and effectively as adjuvant in shingles vaccine Shingrix [53,111,112].

QS-21 is a triterpenoid sapnonin derived from a beta-amyrin scaffold that contains seven different sugar moieties, none of which are glucose (Figure 5.3). The biosynthetic pathway for this compound is unknown, and currently all *Quillaja* saponins for use as food additives and vaccine adjuvants are derived from harvesting of and extraction from tree bark. Work by the Osbourn group (JIC) has begun to sequence the *Q. saponaria* genome and isolate the relevant biosynthetic enzymes.

Given that none of the sugar moieties in QS-21 are glucose, there is an opportunity to utilise the genome data and the recent advances in the understanding of GT1 substrate specificity to mine the genome for potential candidates. Furthermore, it is unknown whether the genes for QS-21 biosynthesis are found in a BGC or are distributed across the genome.

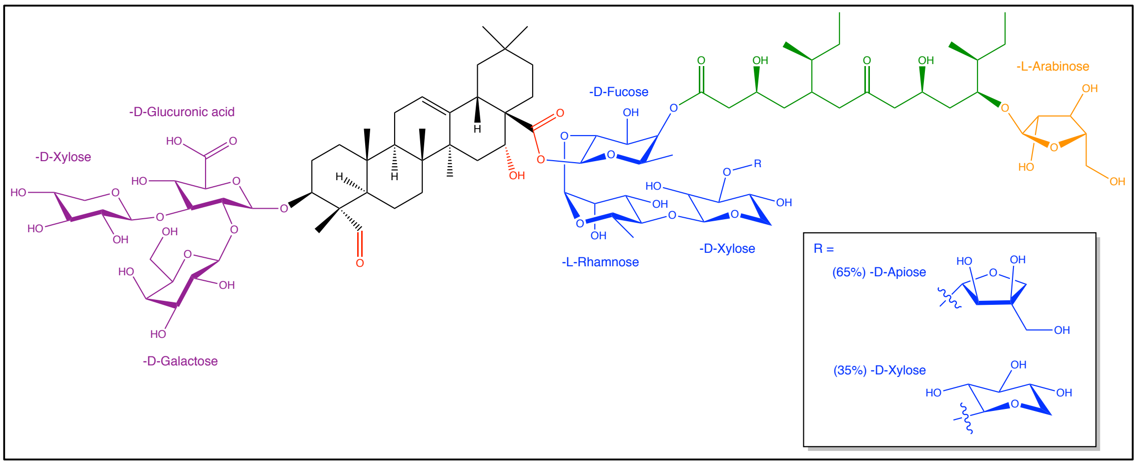


Figure 5.3 **The structure of QS-21**
The beta-amyrin triterpene scaffold is shown in black and products of CYP modifications in red. The branched trisaccharide is shown in purple, the linear tetrasaccharide in blue, the acyl chain in green and the attached arabinose moiety in orange.

5.1.3 Aims

The aims of this chapter are to:

- Incorporate the conclusions found by [30] into a prediction tool for GT1 function
- Apply this tool to putative GT1s from the *Q. saponaria* genome and validate the results against enzyme characterisation work ongoing in the Osbourn group (JIC)
- Inspect the organisation of QS-21 biosynthetic genes in the *Q. saponaria* genome to observe what, if any, relevant BGCs exist

5.2 Methods

5.2.1 GT1 prediction model

A model for predicting the functions of GT1s was developed by incorporating the conclusions made by [30] into an alignment pipeline which identified key residues. Specifically, four sugar-donor specific residues (SSRs) were chosen for prediction of putative GT1 sugar specificity. For UGT71G1 (Figure 5.2) SSR1 is S25, SSR2 is T143 (part of the N5 loop), SSR3 is W360 (W22 of the PSPG motif) and SSR4 is Q382 (Q44 of the PSPG motif). Three catalytic residues (CRs) were also selected for their role in the S_N2-like catalytic mechanism of GT1s. For UGT71G1: CR1 is H22, CR2 is D121 and CR3 is S612. For prediction of acceptor/function specificity, pHMMs were generated from characterised GT1s corresponding to the phylogenetic families reported in Figure 5.1. In some cases, these were broken into subfamilies, such as group L which contains three function-specific monophyletic groups and a fourth group with no conserved function (Figure 5.1). The locations of the SSRs and CRs, in addition to the pHMMs, were then able to be utilised for the annotation of uncharacterised GT1s. Figure 5.4 summarises this methodology.

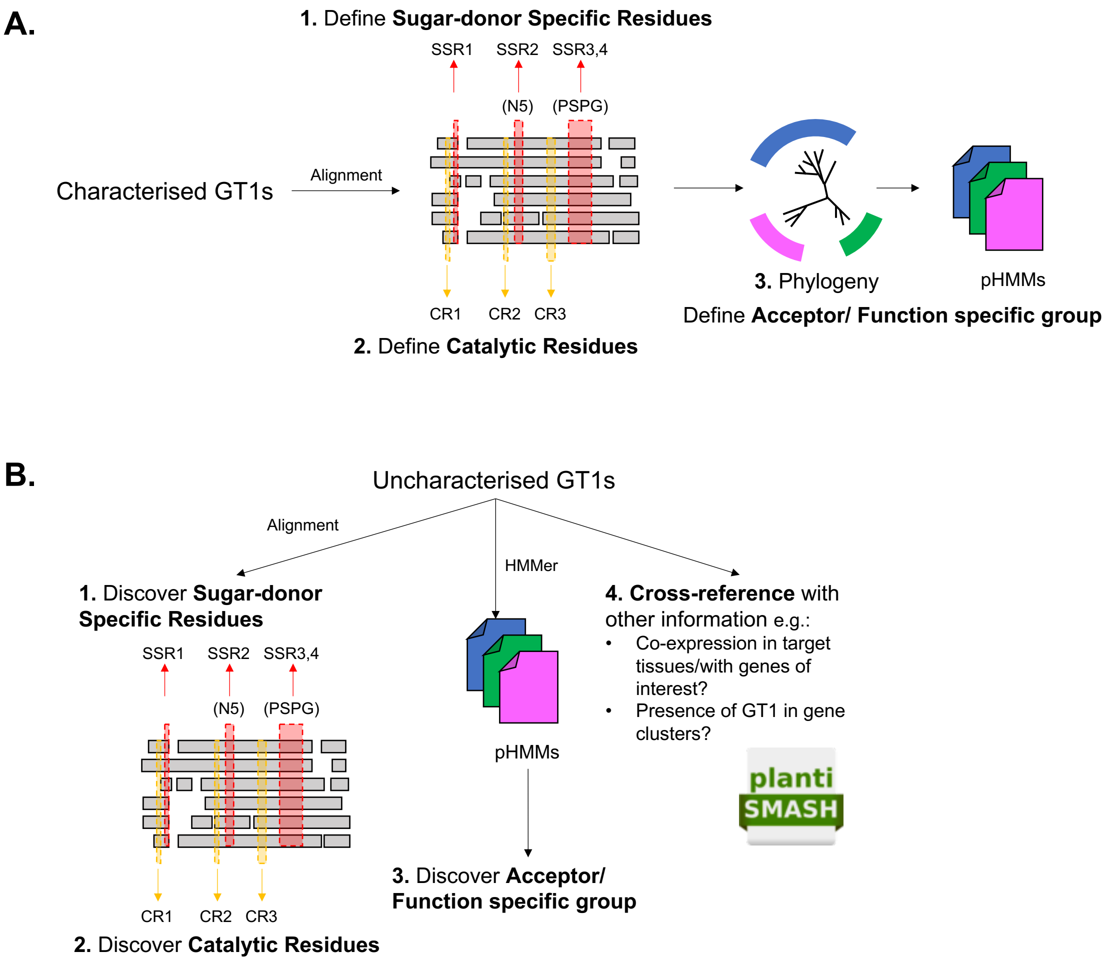

A.
1. Define Sugar-donor Specific Residues
SSR1
SSR2
SSR3,4
(N5)
(PSPG)
Characterised GT1s
Alignment
CR1
CR2
CR3
2. Define Catalytic Residues
3. Phylogeny
pHMMs
Define Acceptor/ Function specific group
B.
Uncharacterised GT1s
Alignment
HMMer
1. Discover Sugar-donor Specific Residues
SSR1
SSR2
SSR3,4
(N5)
(PSPG)
CR1
CR2
CR3
2. Discover Catalytic Residues
pHMMs
3. Discover Acceptor/ Function specific group
4. Cross-reference with other information e.g.:
• Co-expression in target tissues/with genes of interest?
• Presence of GT1 in gene clusters?
plantiSMASH

Figure 5.4 **Computational workflow for A) building and B) applying the pipeline for functional prediction of GT1s**
A) Characterised GT1s (from [30]) were aligned and key residues for sugar-donor specificity (SSRs) and catalytic activity (CRs) extracted as described above. A phylogeny was then generated and the sequences within monophyletic groups defined in Figure 5.1 were used to generate pHMMs. B) Uncharacterised GT1s were aligned and the SSRs and CRs found. HMMer was used to classify the acceptor/function-specific phylogenetic group. These data can then be cross-referenced with other information of potential interest, such as the presence of target GT1s in putative BGCs and/or the assessment of expression data.

5.2.2 Quillaja saponaria analysis

A set of putative GT1s were taken from a draft *Q. saponaria* genome (Osbourn Group, JIC), as defined by alignment to InterPro domain IPR001296. SSRs and CRs were extracted via alignment to the characterised GT1s and HMMer used to find the closest matching acceptor/function specific pHMMs. Pearson correlation coefficient (PCC) values were calculated for GT1 co-expression with the beta-amyrin synthase responsible for making the QS-21 triterpene backbone. The presence of candidates in putative BGCs as identified by plantiSMASH 1.0 (as implemented in Chapter 2) was cross-referenced and noted.

5.3 Results and discussion

5.3.1 Predicting function of Quillaja saponaria GT1 enzymes

Four sugar-donor specific residues (SSRs) were chosen for prediction of putative GT1 sugar specificity, one of which is present in the N5 loop and two of which in the PSPG motif. Three catalytic residues (CRs) were chosen to report whether a canonical S_N2-like reaction mechanism was likely. A profile-based approach was used to classify putative GT1 acceptor/function specificity according to Figure 5.1. This approach is detailed above (5.2.1) and summarised Figure 5.4.

Putative GT1 sequences were taken from a draft *Quillaja saponaria* genome (Osbourn Group (JIC)) and the above classification pipeline applied. Co-expression with the QS-21 biosynthetic pathway genes was quantified and plantiSMASH 1.0 output cross-referenced to determine which, if any, GT1s were found in putative BGCs. A wide range of putative functions were predicted for the GT1s assessed and these data are summarised in Table 5.1.

Table 5.1 **Predictions of *Quillaja saponaria* GT1 function**

Genes in bold have since been characterised as part of the QS-21 biosynthetic pathway. Stars indicate particular candidates of interest for the remaining steps. Co-expression with QS-21 pathway genes is shown by the Pearson correlation coefficient (PCC) of each gene vs the beta-amyrin synthase. Cross-reference with plantiSMASH 1.0 output is also shown, where genes form part of putative BGCs.

			SSR					CR					
	Gene ID	Acceptor/function profile	1	2	3	4	SSR predictions	1	2	3	CR notes	PCC	plantiSMASH BGC
	QUISA32244_Elv1_0321930	**L: Ester-forming**	**P**	**F**	**T**	**Q**	**Novel, Not Glc/Gal**	**H**	**D**	**S**		**0.98**	**BGC9: Saccharide**
★	QUISA32244_Elv1_0213700	D: Triterpenoid_Flavonoid-7-O-	P	V	W	H	Gal/Ara	H	D	S		0.97	
	QUISA32244_Elv1_0123860	**D: Triterpenoid_Flavonoid-7-O-**	**P**	**G**	**W**	**H**	**Gal/Ara**	**H**	**D**	**S**		**0.96**	
	QUISA32244_Elv1_0321920	**A: Glycosidic_branch_elongating**	**P**	**V**	**S**	**Q**		**H**	**D**	**S**		**0.96**	**BGC9: Saccharide**
★	QUISA32244_Elv1_0131010	D: Triterpenoid_Flavonoid-7-O-	P	I	W	Q	Xyl, Not Glc/Gal	H	D	S		0.94	BGC26: Saccharide
	QUISA32244_Elv1_0213710	D: Triterpenoid_Flavonoid-7-O-	P	G	W	Q		H	D	S		0.94	
	QUISA32244_Elv1_0283870	**D: Triterpenoid_Flavonoid-7-O-**	**P**	**V**	**W**	**Q**		**H**	**D**	**S**		**0.93**	**BGC11: Saccharide**
	QUISA32244_Elv1_0082410	A: Glycosidic_branch_elongating	P	V	P	Q		H	D	S		0.92	
	QUISA32244_Elv1_0131000	D: Triterpenoid_Flavonoid-7-O-	-	-	S	H	Gal/Ara	-	-	S		0.92	BGC26: Saccharide
★	QUISA32244_Elv1_0101700	D: Triterpenoid_Flavonoid-7-O-	P	A	W	Q	Novel, Not Glc/Gal	H	D	S		0.91	
★	QUISA32244_Elv1_0234130	D: Triterpenoid_Flavonoid-7-O-	P	P	W	Q	Ara	H	D	S		0.90	BGC38: Saccharide
	QUISA32244_Elv1_0264750	D: Triterpenoid_Flavonoid-7-O-	-	-	W	Q		-	-	S		0.89	BGC32: Saccharide
★	QUISA32244_Elv1_0084600	A: Glycosidic_branch_elongating	P	I	R	H	Xyl, Not Glc/Gal,GlcA,Gal/Ara	H	D	S		0.89	BGC4: Saccharide
	QUISA32244_Elv1_0032650	L: Ester-forming	P	H	-	-	Novel	H	D	S		0.88	
	QUISA32244_Elv1_0213660	D: Triterpenoid_Flavonoid-7-O-	P	S	W	Q	Xyl/GlcA, Not Glc/Gal	H	D	S		0.87	
	QUISA32244_Elv1_0023500	D: Triterpenoid_Flavonoid-7-O-	P	V	-	Q		H	D	S		0.85	
	QUISA32244_Elv1_0038000	D: Triterpenoid_Flavonoid-7-O-	P	G	W	Q		H	D	T		0.84	
	QUISA32244_Elv1_0037940	D: Triterpenoid_Flavonoid-7-O-	-	S	W	Q	Xyl/GlcA, Not Glc/Gal	-	D	T		0.83	
	QUISA32244_Elv1_0032640	L: Ester-forming	P	H	A	Q		H	D	S		0.83	
	QUISA32244_Elv1_0152180	R: Flavonoid-C-	P	T	W	Q		H	D	S		0.82	
	QUISA32244_Elv1_0283860	D: Triterpenoid_Flavonoid-7-O-	-	-	W	Q		-	-	S		0.81	BGC11: Saccharide
	QUISA32244_Elv1_0028380	A: Glycosidic_branch_elongating	S	T	S	Q		H	D	S		0.81	
	QUISA32244_Elv1_0082430	A: Glycosidic_branch_elongating	-	-	P	Q		-	-	-		0.78	
	QUISA32244_Elv1_0283850	D: Triterpenoid_Flavonoid-7-O-	P	V	W	Q		H	D	S		0.77	BGC11: Saccharide
	QUISA32244_Elv1_0023480	D: Triterpenoid_Flavonoid-7-O-	-	-	W	-		-	-	-		0.75	
	QUISA32244_Elv1_0182920	F: Flavonoid-3-O-	P	L	-	-	Novel, Not Glc/Gal,Novel	H	-	-		0.74	
	QUISA32244_Elv1_0022790	M:	P	T	W	Q		H	D	S		0.73	BGC48: Saccharide
	QUISA32244_Elv1_0102040	D: Triterpenoid_Flavonoid-7-O-	P	-	W	Q		H	-	S		0.72	BGC18: Saccharide
	QUISA32244_Elv1_0184200	L: Flavonoid-5-O-	A	V	-	-	Novel	T	S	-	Non-canonical	0.71	
	QUISA32244_Elv1_0037990	D: Triterpenoid_Flavonoid-7-O-	-	-	W	Q		-	-	-		0.71	
	QUISA32244_Elv1_0130970	D: Triterpenoid_Flavonoid-7-O-	P	I	W	Q	Xyl, Not Glc/Gal	H	D	S		0.71	BGC26: Saccharide
	QUISA32244_Elv1_0031010	F: Flavonoid-3-O-	S	-	-	H	Gal/Ara	H	-	T		0.71	
	QUISA32244_Elv1_0219410	G: Monoterpenoid_cyanogenic_glucoside	P	T	W	Q		H	D	S		0.71	
	QUISA32244_Elv1_0084580	A: Glycosidic_branch_elongating	S	T	-	-		H	D	-		0.70	BGC4: Saccharide
	QUISA32244_Elv1_0022800	M:	P	T	W	Q		H	D	S		0.70	BGC48: Saccharide
	QUISA32244_Elv1_0131030	D: Triterpenoid_Flavonoid-7-O-	P	I	W	Q	Xyl, Not Glc/Gal	H	D	S		0.69	BGC26: Saccharide
	QUISA32244_Elv1_0232000	P:	P	T	-	-		H	D	-		0.66	
	QUISA32244_Elv1_0131050	D: Triterpenoid_Flavonoid-7-O-	-	G	W	Q		-	D	S		0.65	BGC26: Saccharide
	QUISA32244_Elv1_0234150	D: Triterpenoid_Flavonoid-7-O-	P	P	W	Q	Ara	H	D	S		0.63	BGC38: Saccharide
	QUISA32244_Elv1_0031760	G: Monoterpenoid_cyanogenic_glucoside	P	T	W	Q		H	D	S		0.62	BGC28: Saccharide
	QUISA32244_Elv1_0199080	G: Monoterpenoid_cyanogenic_glucoside	P	T	W	Q		H	D	S		0.61	
	QUISA32244_Elv1_0127020	R: Flavonoid-C-	P	T	W	Q		H	D	S		0.58	
	QUISA32244_Elv1_0055340	E: Flavonoid-7-O	-	-	W	Q		-	-	S		0.58	
	QUISA32244_Elv1_0234140	D: Triterpenoid_Flavonoid-7-O-	P	P	W	Q	Ara	H	D	S		0.57	BGC38: Saccharide
	QUISA32244_Elv1_0156490	L: Flavonoid-5-O-	P	I	W	Q	Xyl, Not Glc/Gal	H	T	S	Non-canonical	0.56	
	QUISA32244_Elv1_0199070	G: Monoterpenoid_cyanogenic_glucoside	P	T	-	-	Novel	H	D	-		0.52	
	QUISA32244_Elv1_0213690	D: Triterpenoid_Flavonoid-7-O-	P	V	W	H	Gal/Ara	H	D	S		0.48	
	QUISA32244_Elv1_0123910	D: Triterpenoid_Flavonoid-7-O-	-	-	W	Q		-	-	-		0.46	
	QUISA32244_Elv1_0037980	D: Triterpenoid_Flavonoid-7-O-	-	A	W	Q	Novel, Not Glc/Gal	-	-	T		0.45	
	QUISA32244_Elv1_0294720	M:	P	G	W	Q		H	D	S		0.44	
	QUISA32244_Elv1_0091360	M:	P	L	-	Q	Novel, Not Glc/Gal	H	D	S		0.43	BGC1: Saccharide
	QUISA32244_Elv1_0131040	D: Triterpenoid_Flavonoid-7-O-	P	G	W	Q		H	D	S		0.40	BGC26: Saccharide
	QUISA32244_Elv1_0131060	D: Triterpenoid_Flavonoid-7-O-	-	G	-	Q		-	D	-		0.39	BGC26: Saccharide
	QUISA32244_Elv1_0127010	R: Flavonoid-C-	P	T	W	Q		H	D	S		0.38	
	QUISA32244_Elv1_0326980	L: Flavonoid-5-O-	-	V	-	-		-	S	-	Non-canonical	0.36	
	QUISA32244_Elv1_0195760	C:	P	-	W	Q		H	D	S		0.36	
	QUISA32244_Elv1_0031700	G: Monoterpenoid_cyanogenic_glucoside	P	A	-	-	Novel, Not Glc/Gal	R	D	S	Non-canonical	0.36	BGC28: Saccharide
	QUISA32244_Elv1_0213680	D: Triterpenoid_Flavonoid-7-O-	P	-	W	H	Gal/Ara	H	D	S		0.34	
	QUISA32244_Elv1_0130050	O: Cytokinin-O-	Q	S	W	Q	Xyl/GlcA, Not Glc/Gal	H	D	T		0.33	
	QUISA32244_Elv1_0031670	G: Monoterpenoid_cyanogenic_glucoside	-	A	W	Q	Novel, Not Glc/Gal	-	D	S	Non-canonical	0.33	BGC28: Saccharide
	QUISA32244_Elv1_0245140	D: Triterpenoid_Flavonoid-7-O-	P	A	W	Q	Novel, Not Glc/Gal	R	D	S	Non-canonical	0.31	
	QUISA32244_Elv1_0210210	F: Flavonoid-3-O-	-	I	C	H	Xyl, Not Glc/Gal,Gal,Gal/Ara	-	D	T		0.27	
	QUISA32244_Elv1_0130990	D: Triterpenoid_Flavonoid-7-O-	-	G	-	-		-	-	-		0.23	BGC26: Saccharide
	QUISA32244_Elv1_0273540	L: Ester-forming	P	T	W	Q		H	D	S		0.21	
	QUISA32244_Elv1_0192450	A: Glycosidic_branch_elongating	P	I	W	Q	Xyl, Not Glc/Gal	H	D	S		0.21	
	QUISA32244_Elv1_0117760	L: Flavonoid-5-O-	P	I	W	Q	Xyl, Not Glc/Gal	H	N	S	Non-canonical	0.20	
	QUISA32244_Elv1_0032420	G: Monoterpenoid_cyanogenic_glucoside	P	V	W	Q		H	E	S	Non-canonical	0.18	
	QUISA32244_Elv1_0127000	R: Flavonoid-C-	P	T	W	Q		H	D	S		0.17	
	QUISA32244_Elv1_0209630	G: Monoterpenoid_cyanogenic_glucoside	P	A	W	N	Novel, Not Glc/Gal	H	D	-		0.17	
	QUISA32244_Elv1_0037660	L: Hydroxycinnamate	-	V	-	-		-	N	-	Non-canonical	0.17	

Due to the relative complexity and incompleteness of our understanding of GT1 sugar-donor specificity, SSR predictions are not always a one-to-one mapping of residue to function but can instead provide a guide as to the likely and/or unlikely donors used by a given enzyme. Where no comment is made, there is a lack of any differentiating data between the SSRs found and glucose or rhamnose specific GT1s. Cases where the predictions are reported as ‘novel’ reflect when one or more SSR is an amino acid of a type not seen in any characterised GT1 thus far (for example, an SSR4 of neither Q, N nor H). The canonical CRs are H, D and T/S for which an S_N2-like reaction mechanism is proposed [30]. It is known that other residues at these positions can exist in functional GT1s, though it is not known how these enzymes catalyse subsequent sugar transfer.

It is noted that many examples contain no aligned residues for some SSR/CRs. Given that the PSPG motif in particular is a defining feature of GT1s, such sequences warranted closer inspection as to why possible misalignment has occurred. It was found that such examples represent partial/missing sequence annotations, so it can be presumed that, generally, high proportions of missing SSR/CRs represent a poorer quality candidate for subsequent study.

5.3.2 Characterised GT1 enzymes in QS-21 pathway verify predictive ability

Of the GT1s identified in Table 5.1, four have been subsequently characterised as functional in the QS-21 biosynthetic pathway and have been termed UGT-11, UGT-AL, UGT-AA and UGT-Q (Anastasia Orme (JIC), James Reed (JIC)). A summary of these characterised enzymes and the predictive information generated is provided in Figure 5.5.

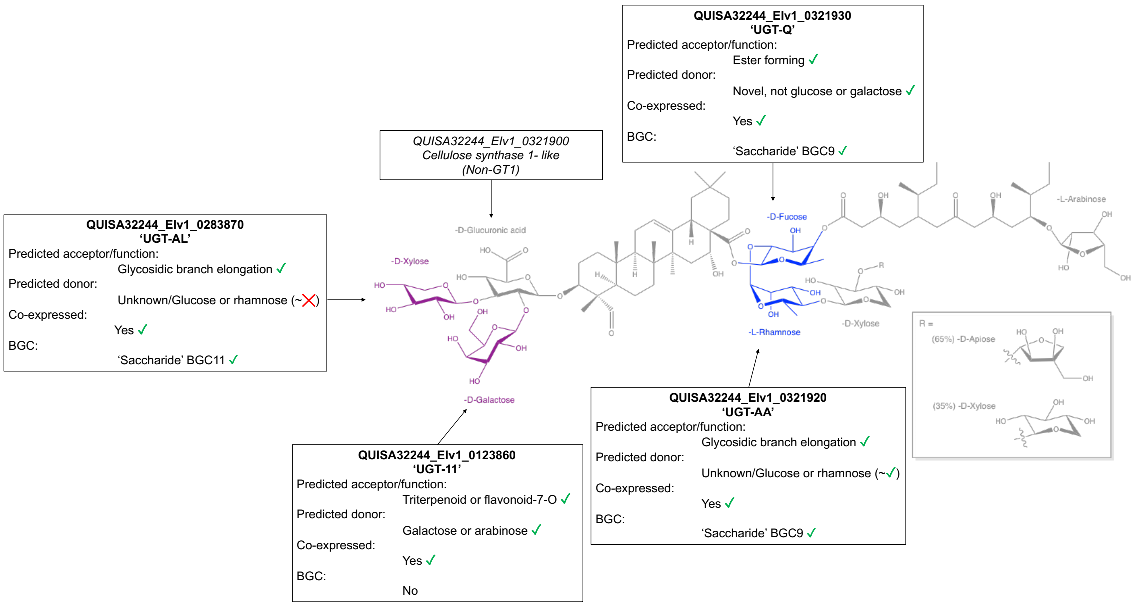


Figure 5.5 **Characterised functions of QS-21 GT1s compared to predictions made**
Four of the five characterised sugar transferase steps in QS-21 biosynthesis indicated above have been attributed to GT1s. For each GT1, a summary of the annotation data generated is given in order to indicate the utility of this process in finding likely GT1s for target glycosylation steps.

The acceptor of all of these GT1s is, naturally, a triterpenoid, and their functions are glycosidic branch elongation (UGT-11, UGT-AL, UGT-AA) and ester formation (UGT-Q). As can be seen, all of the acceptor/function predictions made are consistent with their characterised activities. SSR-based predictions of sugar specificity were successful for three of the four candidates, with only UGT-AL (a xylosyltransferase) not containing any SSRs that would suggest specificity for sugar donors other than UDP-glucose/rhamnose. All of these enzymes were reported with canonical CRs and high co-expression with *QsBAS*. The predictions made are therefore generally consistent the characterised functions, suggesting that this approach has merit for wider application.

Furthermore, there are also a number of potential candidates for GT1s that are predicted to catalyse the xylosylation and arabionsylation steps yet to be characterised in the QS-21 biosynthetic pathway, five of which are marked by stars in Table 5.1. These are all co-expressed, are predicted to act on triterpenoid scaffolds or elongate glycosidic branch chains, have the predicted sugar donor specificity required, contains no missing SSRs or CRs, and three of the five form parts of different putative biosynthetic gene clusters (BGCs 4, 26 and 38). Such results are highly encouraging, and it is hoped characterisation of these enzymes will further reveal the potential for this methodology.

5.3.3 Clustering of QS-21 biosynthetic enzymes

At time of writing, nine genes from *Q. saponaria* have been characterised for QS-21 biosynthesis, including the four GT1s discussed above. An inspection of their position in the genome reveals that some of these genes are found in putative BGCs, whilst others are not. Specifically, the beta-amyrin synthase and two of the three CYPs necessary for the production of the functionalised triterpene scaffold are not co-located with any other biosynthetic genes. UGT-11 is co located with another three GT1s, two which have ~25-30% sequence identity to UGT-11 and one of which is co-expressed. Nine intervening genes upstream of UGT-11 is a co-expressed putative sugar-alcohol dehydrogenase. UGT-AL forms part of the putative ‘BGC 11’ (as defined by plantiSMASH 1.0 analysis) that includes two other co-expressed GT1s and a fatty acid-desaturase. ‘BGC 9’ contains four of the nine QS-21 pathway genes and five other co-expressed putative biosynthetic candidates, including dehydrogenases and BAHD acyltransferases. This is summarised in Figure 5.6.

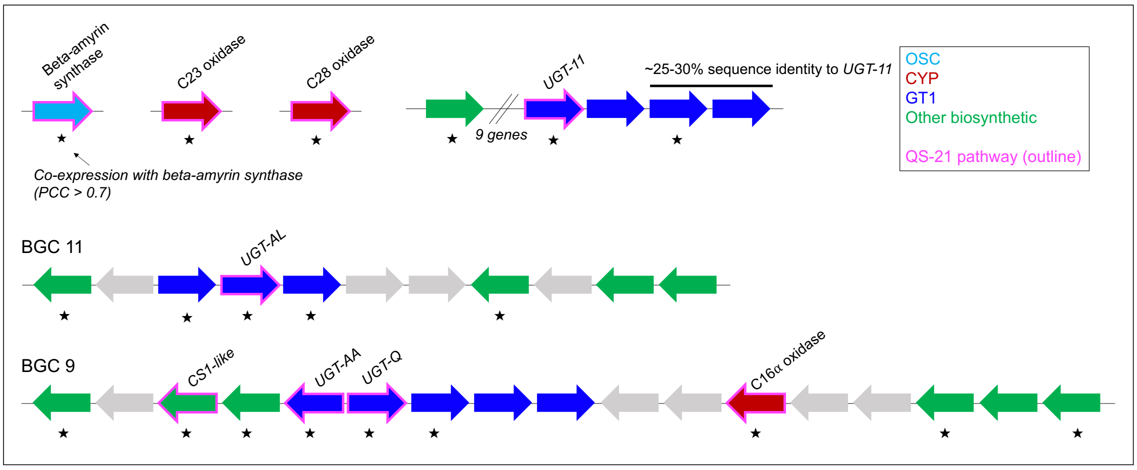


Figure 5.6 **Schematic of partial clustering of QS-21 pathway genes in *Quillaja saponaria***
Nine of the genes required for the QS-21 pathways are distributed across the *Q. saponaria* genome with varying evidence of clustering. Genes for the earlier biosynthetic steps appear to be less likely to be clustered (i.e. OSC, CYPs) in comparison to those for the later stages (e.g. GT1s). Given the large saponin complement of *Q. saponaria*, it is possible that the core triterpene scaffold genes are organised 'generically', with the presence of specific product cluster 'modules' for given compounds.

A total of 58 saponins have so far been isolated from *Q. saponaria* which share a beta-amyrin backbone [108], of which QS-21 is one. It would therefore perhaps be surprising if all of the required biosynthetic genes for the various saponins were to be clustered at a single point in the genome. Evidently, there is not a BGC for the whole QS-21 pathway, but generally the genes responsible for the early steps (i.e. triterpene scaffold synthesis and functionalisation) are not clustered, whilst the genes required for the subsequent scaffold decoration are. The mixed co-localisation observed here suggests that there may be a tendency for specialised glycosylation 'modules' or 'sub-clusters' [32] to form for specific part of the *Quillaja* saponin biosynthetic pathway.

This opens up a pathway for further investigation of this potential phenomenon. The putative biosynthetic steps required to generate the range of saponins isolated from *Q. saponaria* can be compared to the range and distribution of genes encoding for carbohydrate active enzymes across the genome. If there are numerous 'sub-clusters' of such genes, characterisation of their function may show they are restricted to specific branches of the saponin biosynthetic network. This hypothesis is supported by large number of 'saccharide' BGCs defined by plantiSMASH across plant species (Figure 2.2) that do not encode for the substrate on which such enzymes might act upon. As such, characterisation of the various saponin biosynthetic pathways in *Q. saponaria* may lead to the generation of more general hypothesis with regards to the distribution and function of BGCs across wider plants.

5.4 Conclusions

The work presented in this chapter has resulted in the production of a tool for the prediction GT1 function in plants which can be utilised in BGC genome mining pipelines as well as a standalone analysis of any set of putative GT1 sequences. It may also be used to summarise the repertoire of GT1s in a given plant genome, similar to the approaches presented for OSCs in Chapter 3, in order to build a picture of the evolutionary pathways GT1s have taken across plant species. As more enzymes are characterised, it is hoped that a clearer understanding of the sequence-structure-function relationship is GT1s is developed, which can be incorporated into the predictive capabilities described here.

A considerably more statistically complex and general GT1 prediction tool for plants has been published, called 'GT-Predict' which has demonstrated ability in the functional prediction of all GT1s in *A. thaliana* and uses a full protein sequence and phylogenetically naïve clustering, classification and modelling approach [113]. The methodology presented in this chapter approaches the problem of enzyme classification from the 'opposite' end, in simply reflecting the patterns ascertained by expert study and building upwards instead of performing a full, unbiased clustering and classification approach. It is likely that the method here may be more suitable for specific cases such as GT1s active in triterpene biosynthesis though far less suitable for broader GT1 prediction outputs. A comprehensive comparison of the two approaches for analysing GT1s from *Q. saponaria* would be worthwhile in order to understand the strengths and weaknesses of these methodologies.

The data presented here also demonstrates the power in combining co-expression and co-location data with predictive tools for the selection of candidates. The particular case of 'sub-clustering' observed here is especially intriguing given the variation in clustering that has been observed in the characterised plant BGCs thus far. For example, the triterpene BGCs in *A. thaliana* act in a complex metabolic network for the modulation of root microbiota populations, utilising both clustered and non-clustered genes [1]. Similarly, genes encoding for production of cucurbitane triterpenoids in the Cucurbitaceae are partially clustered [16].

It would be naïve to suggest that biosynthetic gene organisation in plants would fit neatly into hard boundaries of 'clustered' and 'unclustered' and indeed a wide range of possible states has already been observed in plants [32,33]. However, the evolutionary mechanisms for the 'birth', 'life' and 'death' of plant BGCs have yet to be revealed, and has been made especially intriguing given the dynamism in gene organisation plants have been recently shown to display for BGC formation and maintenance [35]. As more is discovered of the *Q. saponaria* regulatory network for saponin production, the *in planta* roles such diverse saponins have and the evolutionary pathway taken to develop it, further hypotheses may be generated and tested. In particular, genomic studies and metabolite profiling of related species, subspecies or even *Quillaja* populations, may reveal how such organisation has evolved.

Chapter 6. Expanding the scope of plant BGC mining

6.1 Introduction

The BGC mining tools discussed in Chapter 2 are limited by two key factors. The first is the reliance on structurally annotated genome data (i.e. those with gene models), which is frequently not made publicly available as part of genome data publications. The success of the profile guided gene finding approach detailed in Chapter 3 provides an opportunity to test the limits of this approach in terms of full BGC mining from DNA sequence.

The second limitation is that triterpene BGC mining efforts have thus far relied on pre-defined gene family profiles and/or known enzymatic pathways (Chapter 2). The broad set of genome data collected here allows an unbiased approach to be tested. Unbiased, enrichment-based approaches have proven successful in novel BGC finding for bacterial genome mining, such as via ClusterFinder [114]. The use of OSC gene neighbourhood analysis has been demonstrated to be powerful in elucidation OSC-CYP co-location patterns (Chapter 4) and as such can be extended to a wider set of target gene families.

This is relevant for biosynthetic genes, given the recent characterisations of a glycosyl-hydrolase family 1 trans-glucosidase in the avenacin A-1 pathway [115] and a cellulose synthase-like gene found to encode a glucuronosyl-transferase in the QS-21 pathway (Chapter 5; Anastasia Orme (JIC), James Reed (JIC)), neither of which are included as potentially carbohydrate-active enzymes in plantiSMASH 1.0 [34].

Furthermore, it provides an opportunity for the discovery of ancillary, non-biosynthetic genes such as regulators and transporters, which would not be reported using biosynthetic pHHMs. There is no evident reason why gene clustering should be limited to genes that constitute only a core biosynthetic pathway, and whilst studies using EC classification do provide a broader remit for the definition of 'metabolic' genes (Figure 2.1, [59]), such approaches are nonetheless still limited to explicitly overlapping metabolic pathway reactions.

6.1.1 Aims

The aims of this chapter are therefore to:

- Use the gene finding methods previously optimised to test the potential for mining BGCs from genomes without structural annotations in order to leverage more of the available data
- Apply the *OSC* neighbourhood analyses to as broad as possible set of gene families in order to generate an unbiased picture of triterpene gene co-localisation across the Viridiplantae

6.2 Methods

6.2.1 Unannotated genome BGC mining

To generate high-quality, full-length sequence data for biosynthetic gene families, pHMMs from plantiSMASH 1.0 were used to mine the SwissProt database [116]. These full-length proteins were used with Selenoprofiles [74] according to Chapter 3, in order to generate genomes annotated with putative gene models for the relevant biosynthetic enzyme families. To predict gene models for the intervening genes, Augustus [70] was used, using suitable 'pre-packaged' training data for various plant species. The resulting annotations were then merged and converted to a format suitable for analysis by plantiSMASH 1.0 [34]. This methodology is summarised in Figure 6.1.

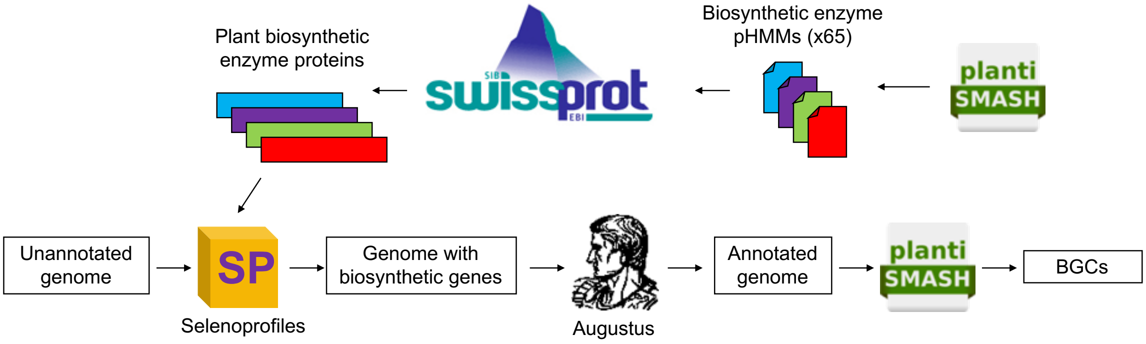

Plant biosynthetic enzyme proteins
Biosynthetic enzyme pHMMs (x65)
Unannotated genome
SP
Selenoprofiles
Genome with biosynthetic genes
Augustus
Annotated genome
BGCs

Figure 6.1 **Summary of methodology for BGC mining of unannotated genomes** Selenoprofiles requires full length protein sequence data for profile-mediated gene finding. To utilise Selenoprofiles for biosynthetic gene-finding, plantiSMASH 1.0 pHMMs were used to extract the corresponding high-quality, full length protein sequences from SwissProt. Augusuts was then used to provide gene models for the rest of the gene. These data were then combined into a full genome annotation for analysis by plantiSMASH 1.0.

6.2.2 Biased and unbiased neighbourhood analysis

For OSC neighbourhood analysis, a similar approach was taken as described in Chapter 4 (Figure 4.2), where OSC flanking genes (+/- 10 genes upstream and downstream) were located and classified using HMMer [62]. For the biased neighbourhood analysis (i.e. restricted to only biosynthetic pHMMs as defined by plantiSMASH 1.0), pHMMs from plantiSMASH 1.0 were used. This approach was then expanded for unbiased neighbourhood analyses, 6675 pHMMs were taken from the Pfam database [63]. These were chosen based on the presence of the Pfam profiles in Viridiplantae according to Pfam taxonomy database [63]. In cases where individual pHMMs exist for N and C terminal domains, gene counts were merged before statistical anaylses with Fisher's exact test. Figure 6.2 summarises the approach used for this unbiased study.

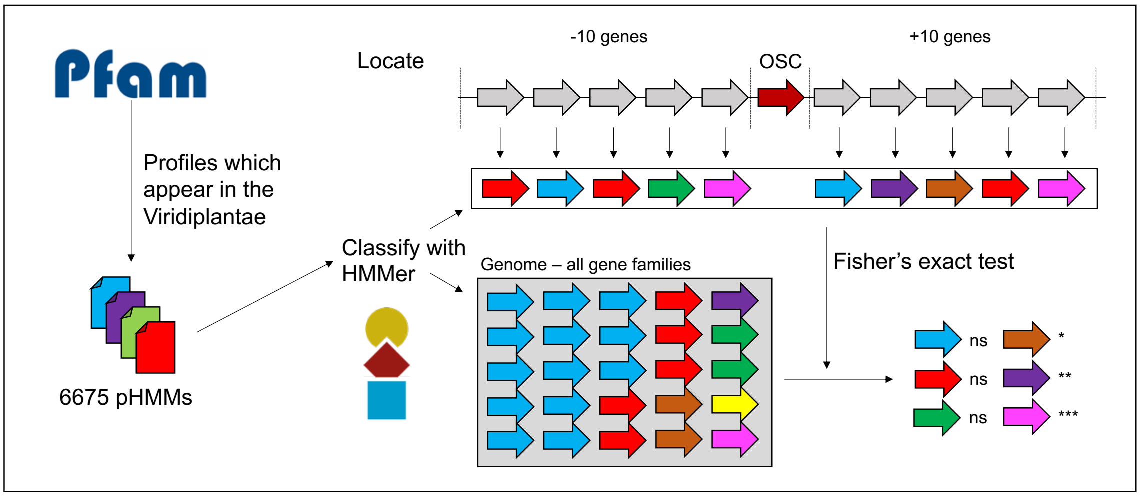


Figure 6.2 **Summary schematic of methodology for unbiased OSC neighbourhood enrichment**
Profiles were filtered from the Pfam database according to occurrence in the Viridiplantae, resulting in 6675 pHMMs for use in plant gene annotation. All genes were assigned to the closest scoring pHMM and neighbourhood enrichment carried out for genes co-located with OSCs as in Chapter 4.

6.3 Results

6.3.1 BGC mining of unannotated genomes

Given that profile-based OSC mining was accurate and automatable (Chapter 3), the prospect of full BGC mining of genome data with no structural annotations (i.e. only DNA sequence) using this approach was assessed. As plantiSMASH uses a density-based parameter for BGC definitions (Chapter 2) an approach which included only biosynthetic profiles would be unsuitable. However, a method which used profile-based alignments for all putative gene families would be highly resource intensive and likely poorly reconstitute full genome annotation pipelines already available [23]. As such, an approach was developed to utilise the Selenoprofiles [74] method for biosynthetic gene families and Augustus, an *ab initio* gene prediction tool [70], for the intervening genes (approach summarised in Figure 6.1). Augustus is distributed with 'pre-packaged' training parameters for specific species, including *A. thaliana* and *Zea mays*.

A comparison of the plantiSMASH outputs for *A. thaliana* using the full reference genome with and without structural annotations is shown in Figure 6.1. Broadly, BGCs are reconstituted accurately, with differences mostly due to BGCs being split or merged across the outputs. This accuracy may be expected, given that Augustus training parameters are well optimised to *A. thaliana* gene structure and distribution.

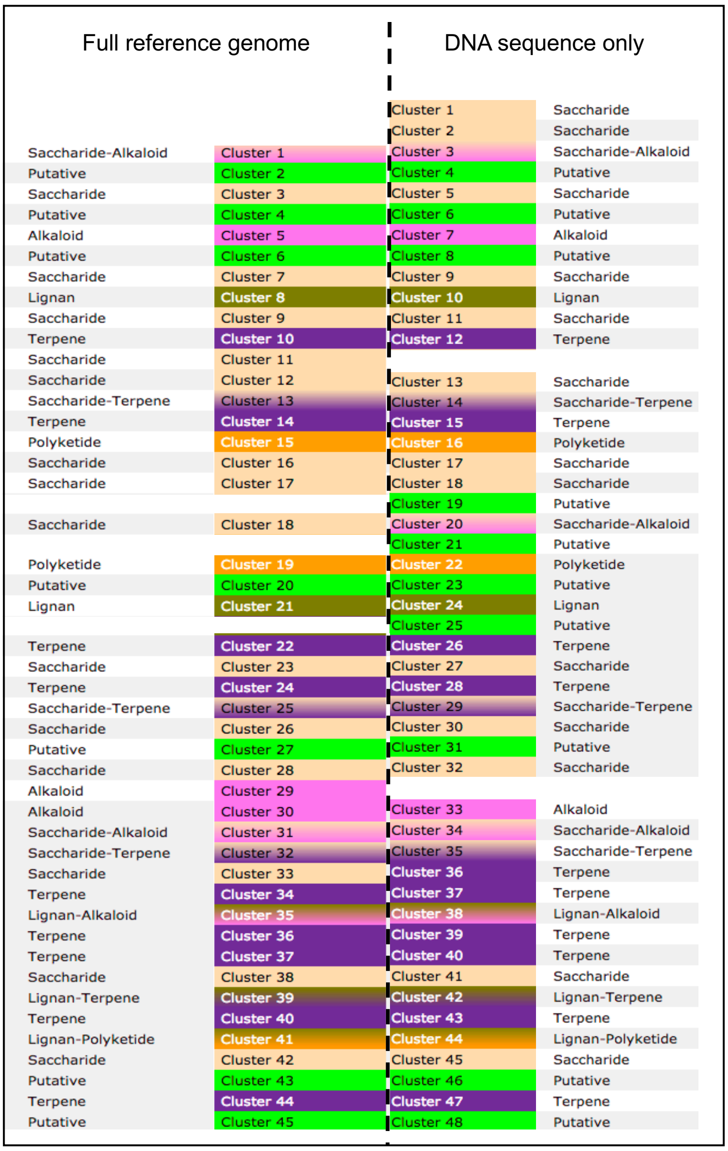


Figure 6.3 **Whole genome BGC analysis from an unannotated genome is successful for *A. thaliana.***

Left: plantiSMASH 1.0 output for the reference *A thaliana* genome. Right: plantiSMASH 1.0 output for a genome based on the DNA sequence of the reference *A. thaliana* genome, with profile-based gene models for biosynthetic genes created by Selenoprofiles [74] and *ab initio* gene models for intervening genes via Augustus [70].

The same method was applied to the *Oryza sativa* Japonica group genome, using the available *A. thaliana* and *Z. mays* training data. Table 6.1 demonstrates the unacceptably high variability and low accuracy in the outputs of this approach. In particular, sensitivity was found to be largely due to a 'malus' parameter required for trimming low quality gene predictions. Perhaps surprisingly, the species from which the training data was derived mattered little, despite *Z. mays* being far more closely related to *O. sativa* than *A. thaliana.* It was therefore concluded that, outside of extremely closely related species for which Augustus training data would be available, approximations of gene density and therefore accurate BGC predictions were unable to obtained using this method.

Table 6.1 **Whole genome BGC analysis from an unannotated *Oryza sativa* genome.** Augustus is unacceptably sensitive to ‘malus’ parameter changes, which appears to impact the accuracy of gene predictions even more so than the species training data chosen. The correctly annotated number of genes in the genome analysed was 39265.

Training species	Augustus malus	BGC counts			Total genes predicted
		True positive	False positive	False negative	
A. thaliana	1	20	22	18	73998
	0.99	17	16	21	66655
	0.98	24	20	14	41602
	0.97	6	36	32	20717
Z. mays	0.99	19	25	19	50440
	0.98	16	32	22	42026
	0.97	15	44	23	35261

This approach was therefore unsuccessful for a use as part of a plantiSMASH mining method, though BGC definitions based on distance would likely be more amenable to this method, assuming the target genes were consistently co-located within a set base-pair range. The novel sesterterpene synthases recently discovered in Brassicaceae [117] are an ideal candidate for this, given they consist of co-located pairs of terpene synthase (TPS) and prenyltransferase (PT) genes. When applied to Brassicaceae genomes for which structural annotations were not available, such as *Capsella bursa-pastoris*, putative sesterterpene synthases were discovered and subsequent analysis has shown them to be functional (data not shown; Ancheng Huang JIC/SUSTech).

6.3.2 OSC neighbour analysis demonstrates co-location of known enzymes.

Given the success of OSC neighbourhood analysis for investigating patterns of CYP co-location (Chapter 4), the same approach was used with all biosynthetic profiles utilised by plantiSMASH. The output for *Brassica oleracea* is given in Figure 6.4 as an example, showing significant co-location of numerous expected gene families such as CYPs, acyltransferases and methyltransferases. As before, Fisher’s exact test was used to determine the significance of gene enrichment. It is argued that, because of the relatively low sample size in the neighbouring gene set and the consequent sensitivity this introduces to minute fluctuations in annotation parameters, the significance values reported should not be treated as strict cut-offs.

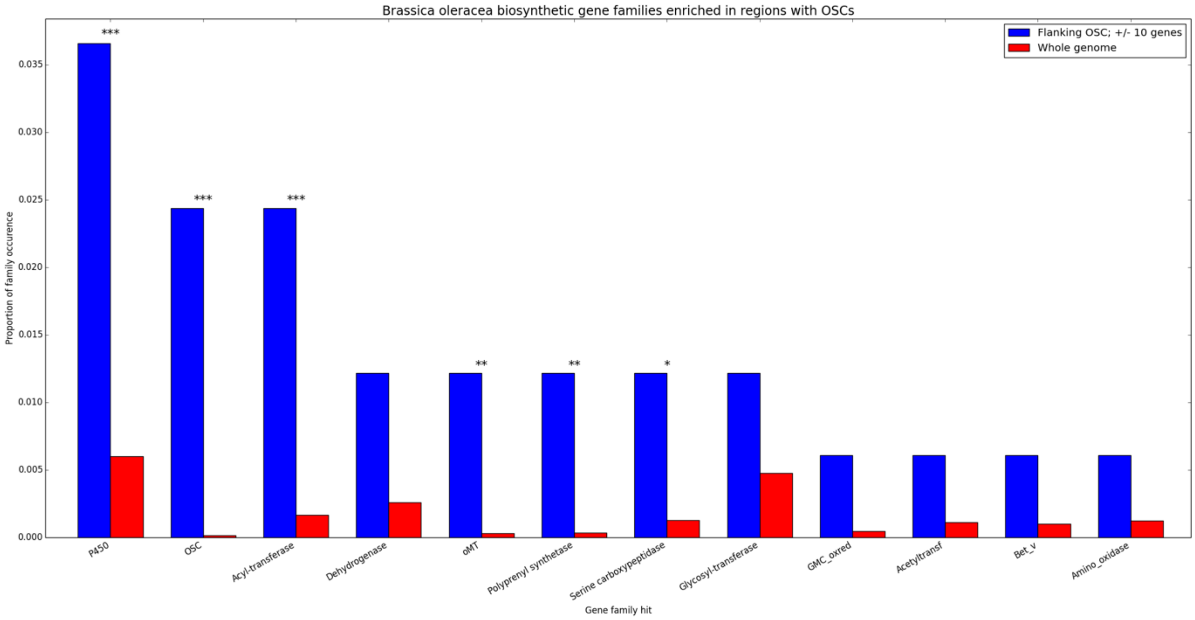


Figure 6.4 **OSC neighbourhood analysis for biosynthetic gene families in *Brassica oleracea***
Bar height represents the relative proportion of given gene families in the OSC neighbourhood compared to the whole genome; *p*-values derived from Fisher's exact test are indicated by asterisks.

Figure 6.5 displays these data across all the species analysed for the eleven most frequently reported gene families. As expected, the most consistently co-located gene family are OSCs themselves, given local gene duplication. In the green algae, basal angiosperms and basal monocots where this is not found, only a single OSC was present in the genome (with the exception of *Asparagus officinalis* (Chapter 3)). Beyond this, the most commonly co-located gene families were those known to be involved in triterpene biosynthetic pathways (i.e. CYPs, acyl-transferases, dehydrogenases, methyl-transferases and glycosyl-transferases).

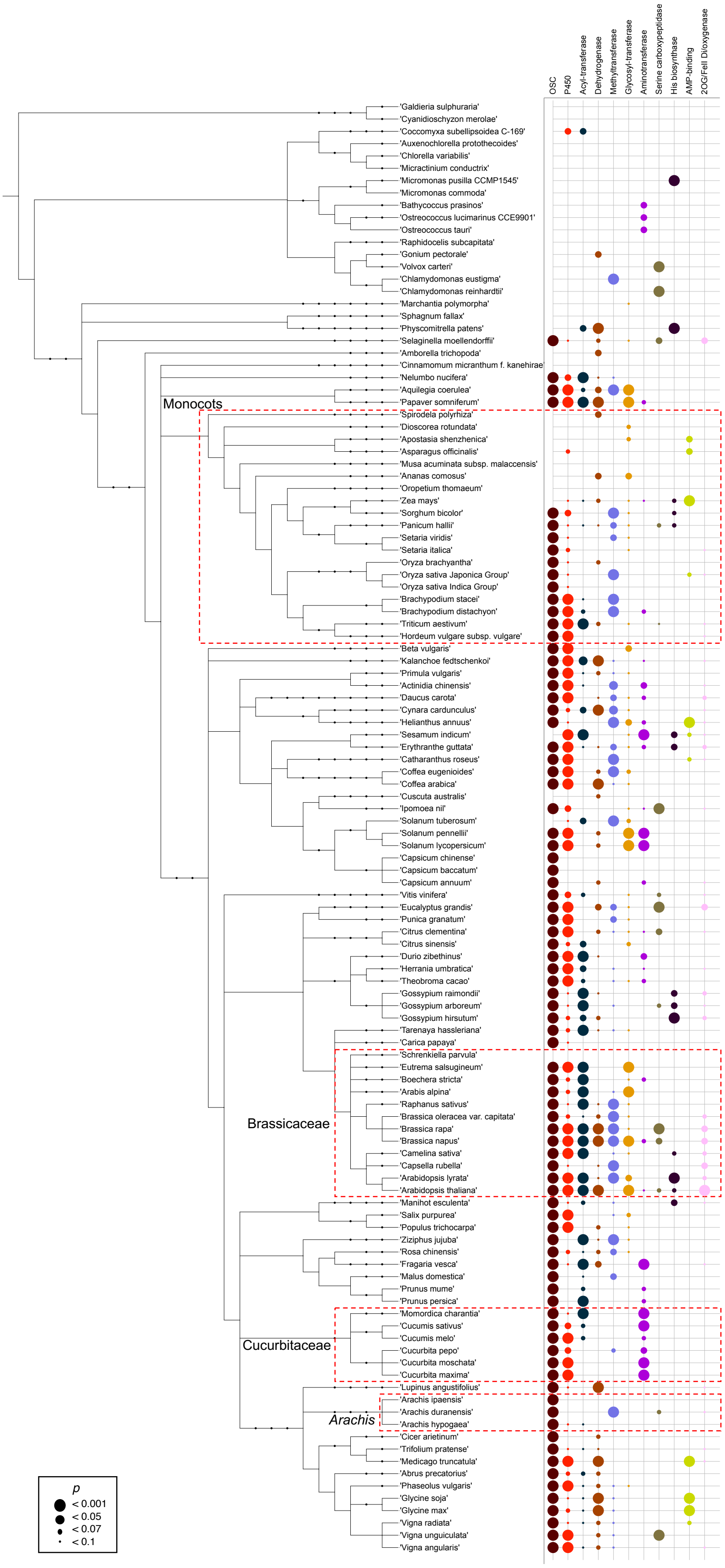


Figure 6.5 **OSC neighbourhood analysis for biosynthetic genes**

Circle size represents significance determined by Fisher’s exact test. Ordered by decreasing frequency of occurrence (left to right). Clades discussed in the text are shown in red boxes.

Comparing the co-location patterns of the Brassicaceae with the Cucurbitaceae provides an example of how these co-location patterns differ between plant clades, consistent with the clade-specific gene organisation previously discussed (Chapters 3, 4). Whilst the significance of acyltransferase co-location is low for *Cucumis* species, it is noted that they do not appear at all in the Cucurbita, indicative of the lack of such enzymes in the cucurbitacin BGCs of those species. The particular co-location of aminotransferases (PF00155) is intriguing, as the role such enzymes may have in triterpene biosynthesis is unclear.

Furthermore, certain dicot species appear to lack any co-location of biosynthetic genes with OSCs, such as the genus *Arachis* which contain a total of 13-14 OSCs present in the individual genomes studied (Figure A1). Yet the genomes of these species do return numerous putative BGCs via plantiSMASH, implying that gene clustering specifically does not occur for triterpene biosynthetic enzymes in these species. As such, they may constitute an interesting case for the study of non-clustered triterpene biosynthesis and regulation to compare to those species with triterpene BGCs.

Overall, these data broadly reconstitute the known triterpene biosynthetic enzyme families and give some indication as to which gene families one may encounter in triterpene BGCs for a given species. These data can be combined with natural product database mining and specific study of the putative BGCs to validate potentially undiscovered pathways.

6.3.3 Unbiased OSC neighbourhood enrichment reveals numerous candidates of interest

Figure 6.2 summarises the approach used for this analysis. A set of 6675 pHMMs was generated for unbiased OSC neighbourhood enrichment, defined by all Pfam profiles found in the Viridiplantae, according to the Pfam taxonomy database [63]. The same enrichment method was applied using these profiles as reported above, using the same genome data. Of the 6675 profiles, 1402 were reported as co-located at least once within ten genes of an *OSC*. Figure 6.6 shows the distribution of p-values (Fisher's exact test) reported for the 30 most frequently reported profiles across all genomes assessed.

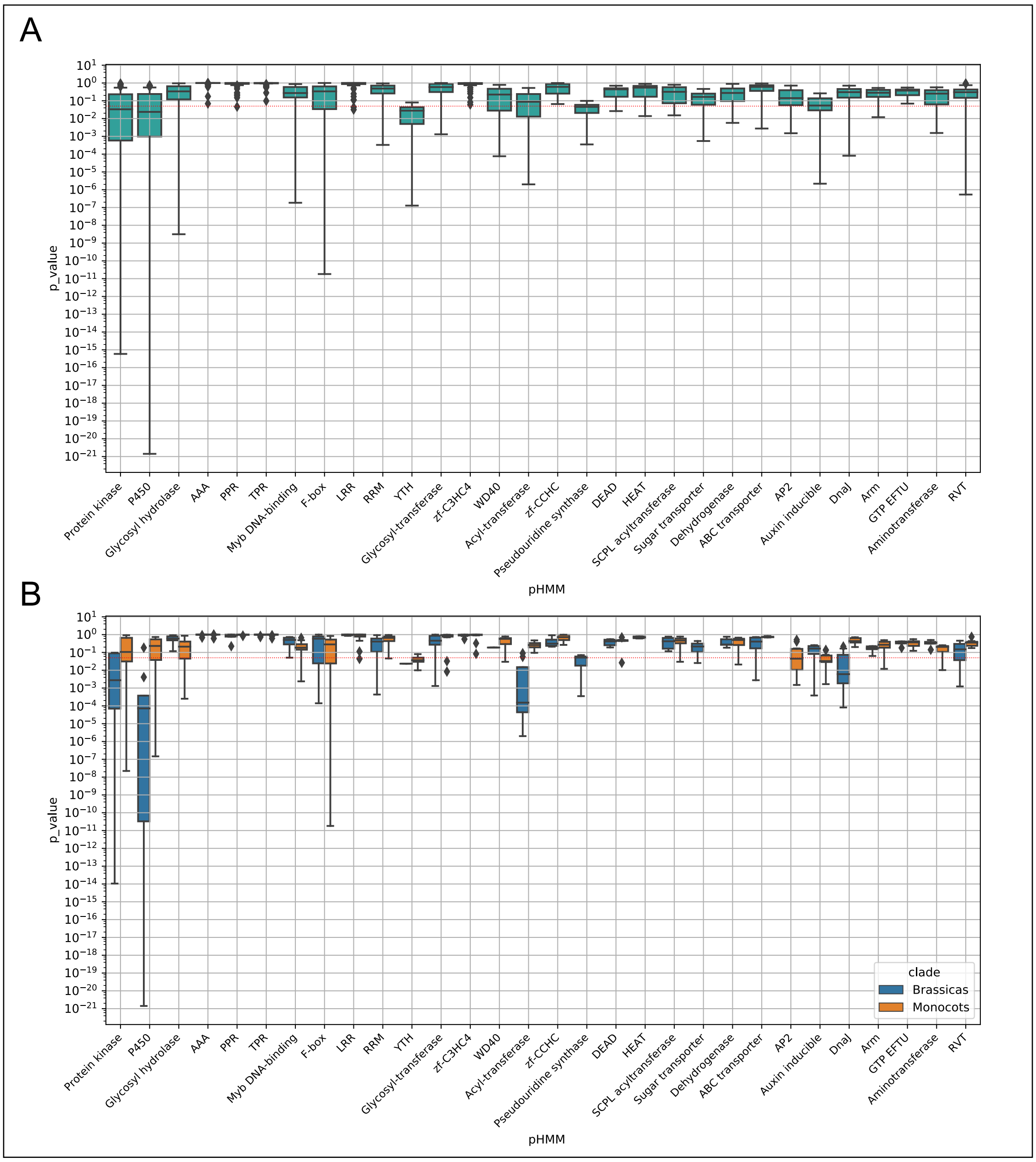


Figure 6.6 **Unbiased OSC neighbourhood analysis**

Showing the 30 most commonly co-located Pfams ordered by frequency of occurrence across all plant species studied, other than other OSCs, with *p*-values (Fisher's exact test) plotted on a log-scale. The red line delineates *p*=0.05. A) Significance across all plant genomes studied. B) Comparison of gene co-location significance between Brassicales and monocot species. Default plotting parameters are used, with the height of the box covering the interquartile range (IQR), and the whiskers extending 1.5x the IQR.

These data firstly highlight the utility of statistical testing to remove uninteresting candidates. For example, ATPases associated with diverse cellular activities (AAA; PF00004) and pentatricopeptide repeats (PPR; PF01535) are often co-located but are not significantly enriched. In Figure 6.6A, the presence of long tails in the *p*-value distributions relative to the mean imply that many gene families are significantly co-located only in a subset of the species analysed. To demonstrate this, Figure 6.6B shows how the significance in enrichment of these gene families can vary between Brassicales and monocot species. For example, the BAHD acyl-transferase family (PF02458) is significantly co-located with OSCs in Brassicales species, but not monocots. Conversely, AP2 transcription factors (PF00847) are not found co-located at all in Brassica species, but are on average more significantly enriched in *OSC* neighbourhood than *CYPs* in monocots.

To more easily interpret these data, median *p*-values and the overall frequency of co-location was taken for each gene family reported and normalised to that of CYP values. Median values were chosen given the strongly biased *p*-value distributions for many of the families reported (note the log scale). These data are plotted and displayed in Figure 6.7.

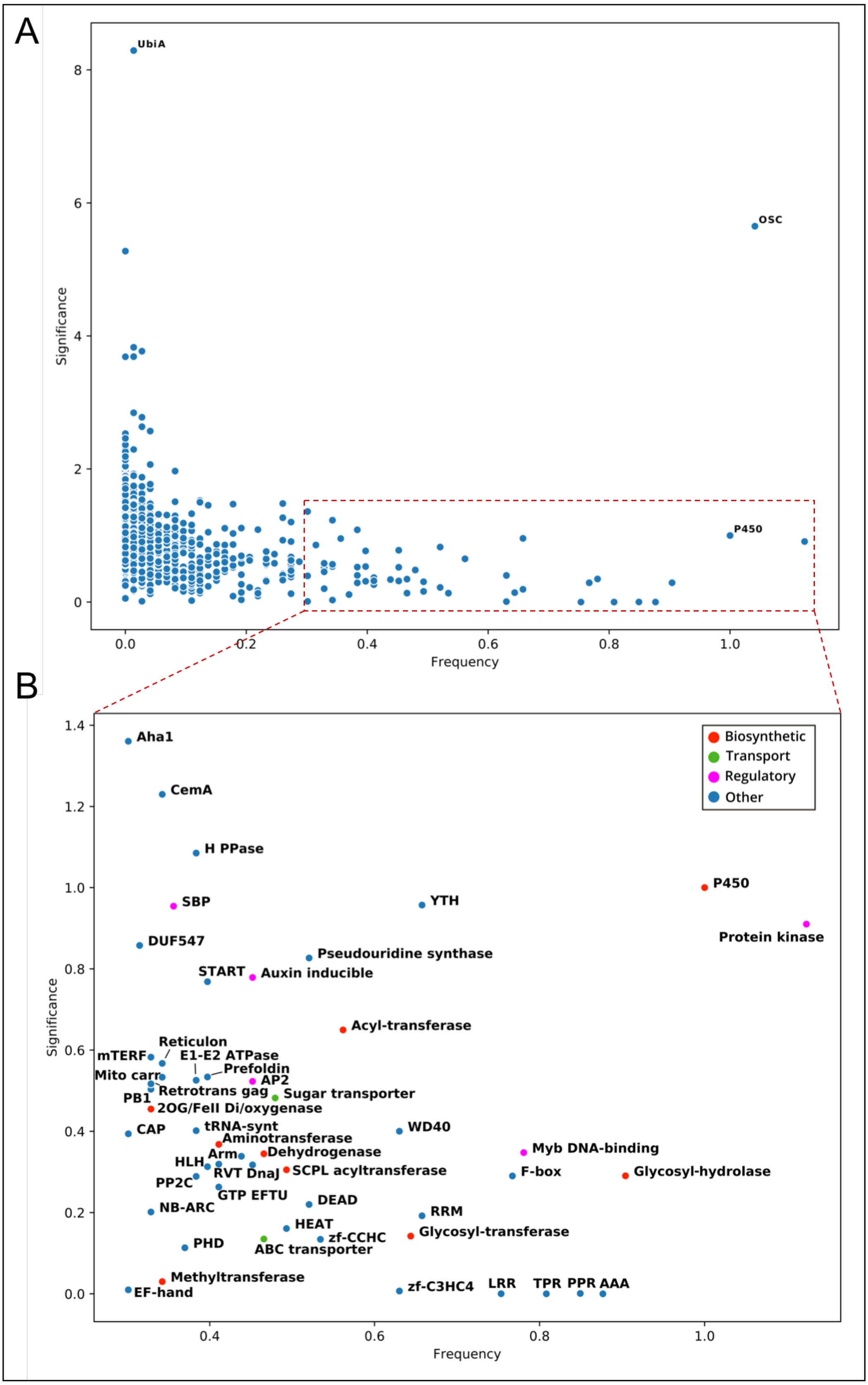


Figure 6.7 **Scatterplot of unbiased OSC co-localisation values**

For each gene family annotated across the genomes studied, the frequency at which they were found co-located with OSC genes was recorded. The significance of this enrichment relative to the overall frequency of that gene family across the genome was calculated by Fisher's exact test. Median *p*-values and overall frequency counts of co-located gene families across all plant genomes studied were normalised to the values found for CYPs (1,1), a gene family known to be significantly co-located with OSCs.

Gene families which are highly significant but only in a very restricted subset of species will tend towards the top-left of Figure 6.7A. For example, UbiA (PF01040) is found co-located with OSCs only in *Daucus carota*. This family has been characterised as containing non-canonical polyprenyl/diterpene synthases in bacteria and fungi with structural homology to type 1 terpene synthases [118] and so may prove to be a candidate of interest for further study in this species.

Figure 6.7B shows the gene families with a frequency of OSC co-location at least 30% of that of CYPs (i.e. ≥ 0.3). Numerous families of interest are evident, including many of the known biosynthetic enzymes previously discussed. Glycosyl-hydrolase family enzymes are found via this method, implying that there may be more TGs to be discovered in triterpene BGCs. However, cellulose synthases do not appear. It is likely that a stricter approach is needed to delineate those enzyme families specifically involved in specialised metabolism given the broad functional scope of cellulose synthase enzymes in plants [119]. This can be achieved through phylogenetic analysis of the cellulose synthase-like enzymes which have been functionally characterised. Of the rest of the biosynthetic gene families reported, the only two that have not been reported as part of triterpene biosynthetic pathways are aminotransferases and 2OG-Fe(II) oxygenases (PF03170), though this family does constitute part of the DIMBOA pathway (Chapter 1, [120]).

Two transporter families are reported, ABC transporters (PF00005) and a subfamily of the major facilitator transporters most frequently found to be involved in sugar transport (PF00083). The inclusion of these is intriguing, given the success of the *Nicotiana benthamiana* transient expression system for elucidating triterpene biosynthetic pathways [21] implies generic transport mechanisms are at least sufficient for the production of the target compounds. Of course, pathways which rely on specific transporters are less likely to have been successfully characterised in any heterologous expression system, and specific *in planta* control of metabolite transport has a wide scope for complexity. These data therefore provide a clear opportunity for further study of putative BGCs containing transporter family genes and the roles they might play.

The inclusion of numerous regulatory gene families is particularly encouraging, given there has been much greater success in characterising the roles of triterpene biosynthetic enzymes than in elucidating their regulatory frameworks. The transcription factor/DNA binding domains Myb (PF00249), SBP (PF03110) and AP2 (PF00847) are reported here specifically. There has been some progress in identifying possible transcriptional regulators for triterpene BGCs. For example, the *Sad1* promoter element of the *A. strigosa* avenacin BGC confers root specific expression across a wide range of higher plant species, wherein a HD-ZIP IV family transcription factor is implicated [38]. For the cucurbitacin BGCs, a basic helix-loop-helix transcription factor is required for gene expression [99]. However, no transcriptions factors have been found as part of any plant BGC thus far [33], making the data presented here particularly interesting for further study. The inclusion of a family of auxin inducible genes (PF02519) and AP2, which is a family characterised by being ethylene responsive, is noteworthy, given the frequent role of triterpenes in plant defence. Finally, the protein kinase family most commonly occurring within the family reported here was pollen receptor-like kinase 1 (PF00069), which are a large trans-membrane kinase family commonly involved in plant development and defence responses [121].

6.4 Conclusions

The reliance on annotated genome data and on pre-defined BGC definitions were identified as two areas in which plant BGC mining tools might be improved (Chapter 2). This chapter summarises approaches to solve these limitations which were achieved with varying success. Whilst a full reconstitution of a genome annotation pipeline was unsuitable for the scope of this thesis, the limits of profile-based and *ab initio* gene prediction in plants were shown, particularly for density-based BGC definitions. Distance-based metrics are naturally easier to implement, but evidently care must be taken to choose suitable parameters specific to the target species and genes.

OSC neighbourhood enrichment has been proven to be a useful tool, not only in demonstrating the ways in which plant clades vary in the repertoire of genes they might use to form BGCs (or indeed highlight the apparent lack of triterpene BGCs) but in identifying novel gene families that have so far been occluded from study. This chapter identifies many avenues for future work in elucidating alternate triterpene biosynthetic pathways and possible candidates for the study of their broader regulation. It also demonstrates that gene clustering may not be limited only to biosynthetic gene families, but may be a universal mechanism for co-ordinated gene regulation.

The gene families discussed here are representative of the broad patterns observed in this study and a wide range of options exist to refine and improve this approach. Firstly, as has been observed, plant clades can utilise different complements of co-located enzyme families and subfamilies in triterpene biosynthesis. Further families of interest that have yet to be studied may therefore be found by separating these data taxonomically. Furthermore, whilst glycosyl-hydrolases and glycosyl-transferases have been found using this unbiased approach, closer inspection of putative BGCs containing carbohydrate active enzymes using Pfam profiles have been inconsistent (data not shown). The CAZy (Carbohydrate Active enZyme) database [122] is a more refined resource for these enzyme families, and the dbCAN2 database provides pHMMs derived from this [123], which is likely to result in a higher quality output if incorporated into this approach.

A further criticism of this approach is the use of Fisher's exact test, which is noted for its relative lack of power [124] that has also been observed here. Whilst gene set enrichment is a non-trivial problem, given the propensity for genes to fall into multiple families, other enrichment statistics exist which have been shown to outperform the classical hypergeometric test by a wide margin [124,125]. Finally, the inclusion of co-expression data would greatly increase the ability of this approach to locate putatively co-regulated gene families and be especially suited to removing false positives from the dataset.

Nonetheless, these data provide the first broad analysis of OSC co-located genes across the Viridiplantae, and provide numerous opportunities for further study of specific and novel gene families in putative BGCs, which can then be included in future genome mining tools. When combined with the advances made in classification of specific gene families (Chapters 3, 4 and 5) these data represent a full and comprehensive analysis of clustered triterpene biosynthetic genes.

Chapter 7. MITE-like sequences in the avenacin BGC

7.1 Introduction

Avena strigosa contains a BGC which encodes the steps for the biosynthesis of avenacin, a root specifc anti-fungal saponin that protects oat species against *Gaeumannomyces graminis* var. *tritici* (commonly known as 'take-all disease') [12,42,115]. The full BGC contains 14 co-expressed genes from six different gene families across a span of 961 kbp (Figure 7.1). Furthermore, the promoter of the *Sad1* gene has been demonstrated to confer root specific expression across a wide range of plant species [38]. However, despite evident co-regulation and tissue-specific transcriptional control, little homology has been discovered between the promotor elements of the clustered genes. One shared homologous region has been found, spanning approximately 270 base pairs that is located approximately 550bp upstream of five genes.

These five genes are all from different gene families (*Sad1*, an OSC; *Sad2*, a CYP; *Sad7*, an acyl-transferase; *Sad9*, a methyl-transferase; and *UGT74H7*, a glycosyltransferase), suggesting that this sequence has been recruited to this position during or after BGC formation. A previous analysis of the sequences indicated they were miniature inverted–repeat transposable elements (MITEs) (Anne Osbourn (JIC)). MITEs are class II non-autonomous transposable elements that do not encode their own transposases [126].

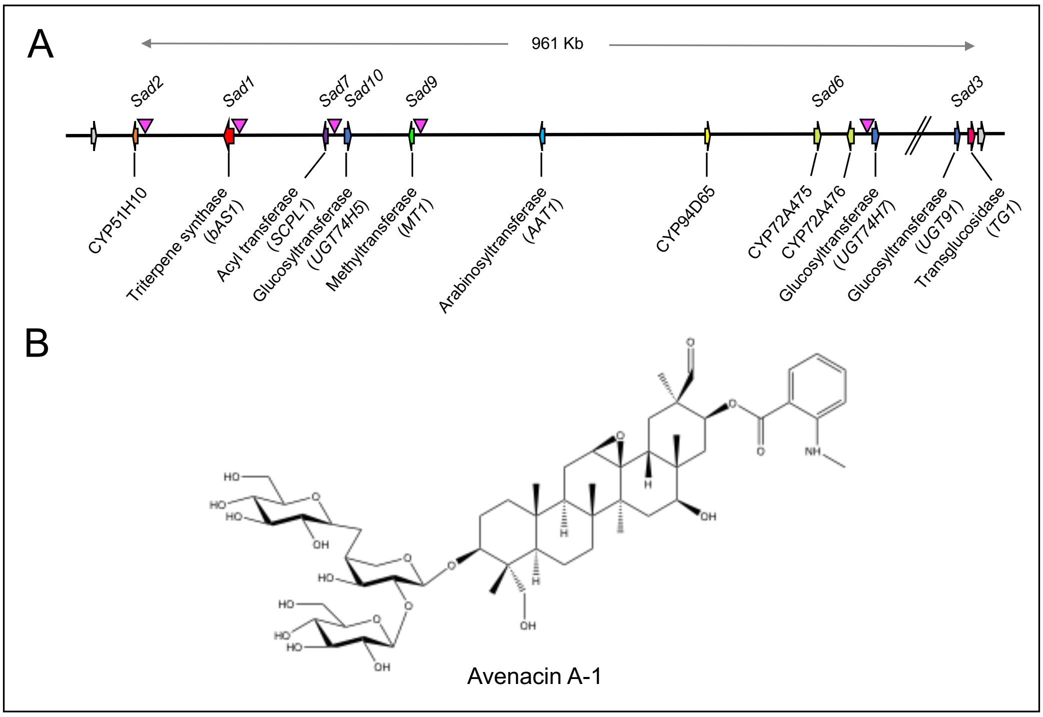


Figure 7.1 **The avenacin BGC in *Avena strigosa***
A) The BGC containing 14 co-expressed genes from six different gene families. MITE-like sequences indicated by pink inverted triangles. B) The product of the BGC avenacin A-1, a saponin which confers fungal resistance in *A. strigosa* roots.

Given the evolutionary role transposable elements can play in the formation of new genes and the rearrangement of gene organisation [127], the possibility that transposable elements may be involved in the creation and/or regulation of BGCs is intriguing. MITEs have previously been specifically implicated in the creation or maintenance of terpene synthase-*CYP* gene pairs in eudicots over other transposable elements such as retrotransposons [37]. Because of the striking conservation of these 'MITE-like sequences' in the avenacin cluster, an opportunity exits to investigate the distribution of homologous sequences in the genome and observe if they may be used to indicate similarly expressed genes, are correlated with genes specific to plant defence and/or are found in other BGCs.

7.1.1 Aims

The aims of this chapter are to:

- Find homologous MITE-like sequences in the *A. strigosa* genome and observe their distribution relative to other genes
- Investigate the expression specificity of genes with homologous MITE-like sequences in the promoter regions
- Test to see if these elements are biased towards being present other putative BGCs in *A. strigosa*, or genes with a similar role in plant defence as avenacin.

7.2 Methods

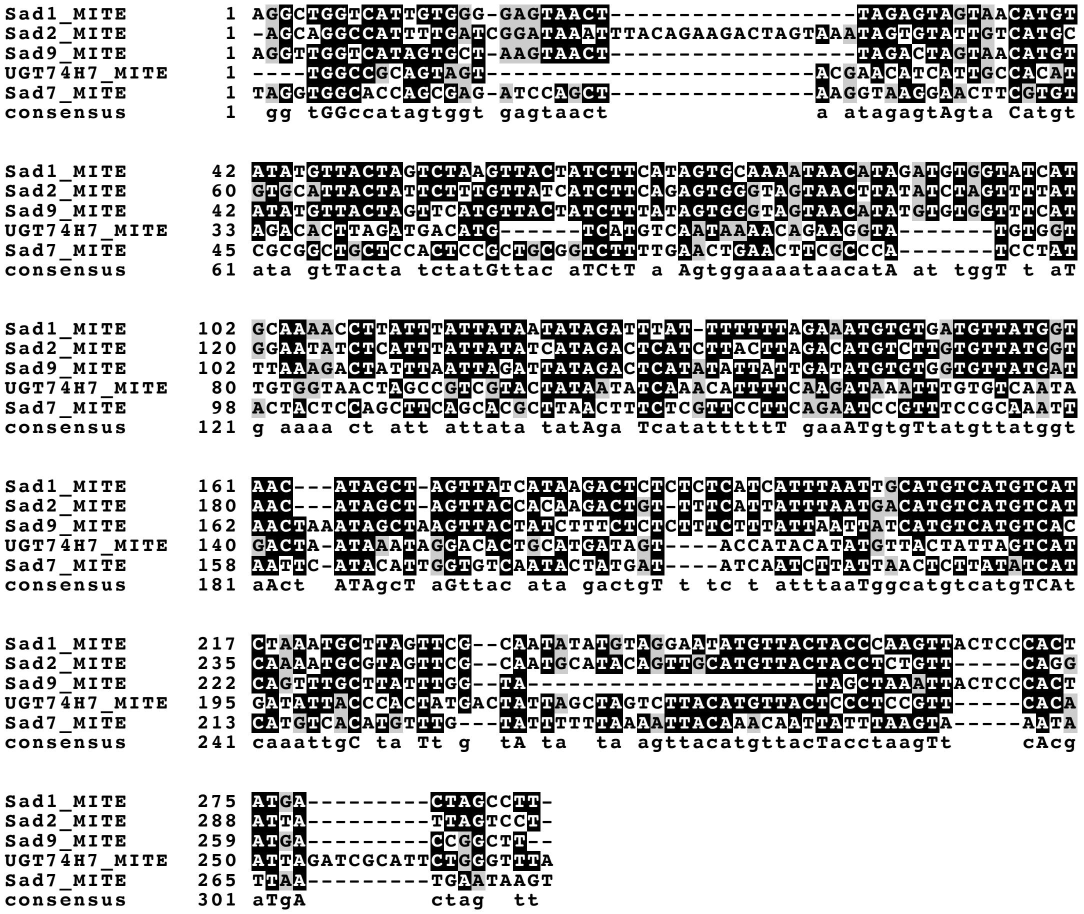
A. strigosa genome and transcriptome data, and outputs from a MITE-Hunter [128] analysis, were provided by Anne Osbourn (JIC) and Bin Han (NCGR CAS). Sequences sharing homology to the MITE family of interest were collected by pHMMer across the genome, using a profile generated from the sequences previously identified in the avenacin BGC (Anne Osbourn (JIC)). Co-expression analysis was carried out using the 'kohonen' [129] package and the 'topGO' [130] package in R [131] was used to assess gene family enrichment for genes with MITE-like sequences present in putative promoter regions (defined by up to 2kbp upstream of the start codon).

A random distribution of MITE-like sequences was simulated using their true gene co-ordinates and randomly assigning them to new co-ordinates on the contigs of the genome. Sequences were prevented from overlapping with each other (or with any exons) as this form of bias in distribution was not under investigation. This distribution of sequences was compared to the actual location of these sequences in the genome to see if any biases were present. Categories were defined as: 5' extended region (2kb - 10kb upstream of ATG), 5' region (0kb - 2kb upstream of ATG), within an exon, within an intron, 3' region (2kb downstream of stop) or intergenic. Homologous sequences were aligned with MAFFT [75] and a phylogeny generated by RAxML [76] using default parameters.

7.3 Results

7.3.1 MITE-like sequences are biased to gene promoter regions

The MITE-like sequences found in the avenacin BGC were judged to be derived from MITEs, but noted as ancient, having undergone considerable sequence turnover (Sue Wessler (UCR)). An alignment of the MITE-like sequences from the avenacin BGC is shown in Figure 7.2. MITEs are known to have a bias in their genomic distribution, being more likely to be found in proximity to genes and specifically promoter elements, therefore an analysis of homologous sequences to the elements in the avenacin BGC across the *A. strigosa* genome was carried out to determine their distribution patterns.



```
Sad1_MITE        1 AGGCTGGTCATTGTGGG-GAGTAACT------------------TAGAGTAGTAACATGT
Sad2_MITE        1 -AGCAGGCCATTTTGATCGGATAAATTTACAGAAGACTAGTAAATAGTGTATTGTCATGC
Sad9_MITE        1 AGGTTGGTCATAGTGCT-AAGTAACT------------------TAGACTAGTAACATGT
UGT74H7_MITE     1 ----TGGCCGCAGTAGT------------------------ACGAACATCATTGCCACAT
Sad7_MITE        1 TAGGTGGCACCAGCGAG-ATCCAGCT---------------AAGGTAAGGAACTTCGTGT
consensus        1  gg tGGccatagtggt gagtaact               a atagagtAgta Catgt

Sad1_MITE       42 ATATGTTACTAGTCTAAGTTACTATCTTCATAGTGCAAAATAACATAGATGTGGTATCAT
Sad2_MITE       60 GTGCATTACTATTCTTTGTTATCATCTTCAGAGTGGGTAGTAACTTATATCTAGTTTTAT
Sad9_MITE       42 ATATGTTACTAGTTCATGTTACTATCTTTATAGTGGGTAGTAACATATGTGTGGTTTCAT
UGT74H7_MITE    33 AGACACTTAGATGACATG------TCATGTCAATAAAACAGAAGGTA-------TGTGGT
Sad7_MITE       45 CGCGGCTGCTCCACTCCGCTGCGGTCTTTTGAACTGAACTTCGCCCA-------TCCTAT
consensus       61 ata gtTacta tctatGttac aTCtT a AgtggaaaataacatA at tggT t aT

Sad1_MITE      102 GCAAAACCTTATTTATTATAATATAGATTTAT-TTTTTTAGAAATGTGTGATGTTATGGT
Sad2_MITE      120 GGAATATCTCATTTATTATATCATAGACTCATCTTACTTAGACATGTCTTGTGTTATGGT
Sad9_MITE      102 TTAAAGACTATTTAATTAGATTATAGACTCATATATTATTGATATGTGTGGTGTTATGAT
UGT74H7_MITE    80 TGTGGTAACTAGCCGTCGTACTATAATATCAAACATTTTCAAGATAAATTTGTGTCAATA
Sad7_MITE       98 ACTACTCCAGCTTCAGCACGCTTAACTTTCTCGTTCCTTCAGAATCCGTTTCCGCAAATT
consensus      121 g aaaa ct att attata tatAga TcatatttttT gaaATgtgTtatgttatggt

Sad1_MITE      161 AAC---ATAGCT-AGTTATCATAAGACTCTCTCTCATCATTTAATTGCATGTCATGTCAT
Sad2_MITE      180 AAC---ATAGCT-AGTTACCACAAGACTGT-TTTCATTATTTAATGACATGTCATGTCAT
Sad9_MITE      162 AACTAAATAGCTAAGTTACTATCTTTCTCTCTTTCTTTATTAATTATCATGTCATGTCAC
UGT74H7_MITE   140 GACTA-ATAAATAGGACACTGCATGATAGT----ACCATACATATGTTACTATTAGTCAT
Sad7_MITE      158 AATTC-ATACATTGGTGTCAATACTATGAT----ATCAATCTTATTAACTCTTATATCAT
consensus      181 aAct  ATAgcT aGttac ata gactgT t tc t atttaaTggcatgtcatgTCAt

Sad1_MITE      217 CTAAATGCTTAGTTCG--CAATATATGTAGGAATATGTTACTACCCAAGTTACTCCCACT
Sad2_MITE      235 CAAAATGCGTAGTTCG--CAATGCATACAGTTGCATGTTACTACCTCTGTT-----CAGG
Sad9_MITE      222 CAGTTTGCTTATTTGG--TA---------------------TAGCTAAATTACTCCCACT
UGT74H7_MITE   195 GATATTACCCACTATGACTATTAGCTAGTCTTACATGTTACTCCCTCCGTT-----CACA
Sad7_MITE      213 CATGTCACATGTTTG---TATTTTTTAAAATTACAAACAATTATTTAAGTA-----AATA
consensus      241 caaattgC ta Tt g  tA ta  ta agttacatgttacTacctaagTt     cAcg

Sad1_MITE      275 ATGA---------CTAGCCTT-
Sad2_MITE      288 ATTA---------TTAGTCCT-
Sad9_MITE      259 ATGA---------CCGGCTT--
UGT74H7_MITE   250 ATTAGATCGCATTCTGGGTTTA
Sad7_MITE      265 TTAA---------TGAATAAGT
consensus      301 aTgA         ctag  tt
```

Figure 7.2 **Alignment of MITE-like sequences from the avenacin BGC**
Alignment created using ESPript [132].

Sequences sharing homology to the MITE sequence were found throughout the *A. strigosa* genome via nHMMer, with a profile constructed of the five sequences found in the avenacin BGC. A phylogenetic analysis of these sequences demonstrated that the elements found in the BGC are not from a conserved phylogenetic group relative to the homologous sequences located elsewhere in the genome (Figure 7.3).

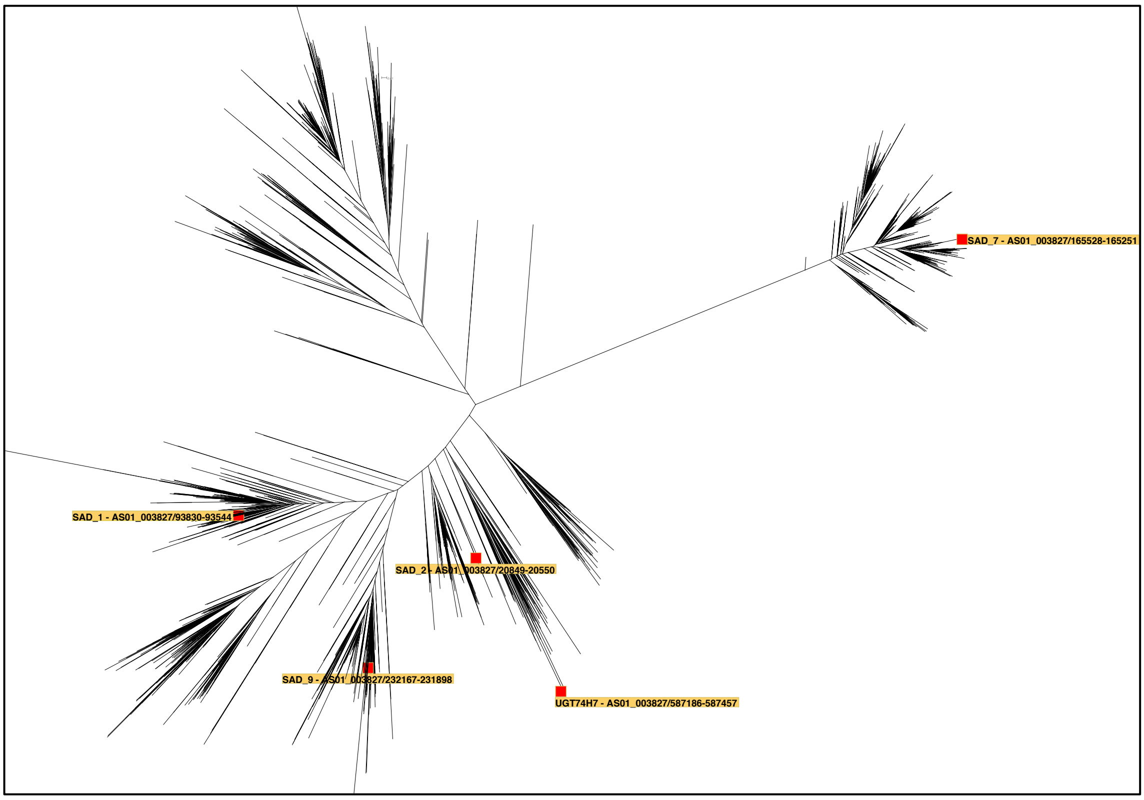

SAD_7 - AS01_003827/165528-165251
SAD_1 - AS01_003827/93830-93544
SAD_2 - AS01_003827/20849-20550
SAD_9 - AS01_003827/232167-231898
UGT74H7 - AS01_003827/587186-587457

Figure 7.3 **Maximum-likelihood phylogeny of MITE-like sequences in *Avena strigosa***
Homologous sequences to MITE-like elements found in the avenacin BGC (labelled) were aligned with MAFFT [75] and a phylogeny generated by RAxML [76] using default parameters.

The spatial distribution of these sequences in relation to genes was observed for the MITE-like sequences across the *A. strigosa* genome and compared to a modelled randomised distribution (Figure 7.4A). No bias was observed for these sequences being on the same or opposite strands of the genes they were in proximity to, nor were any elements with the reverse sequence found.

The distribution of these elements in putative promoter regions of genes was investigated further, given the conserved location in the promoters of the avenacin BGC. A distribution bias was observed, with the most enriched regions for these elements being in the same range as observed in the avenacin BGC (~550 bp), though overall there was a particularly strong presence within the first 2 kbp upstream of the associated gene's start site (Figure 7.4B). These MITE-like sequences are therefore distributed in the *A. strigosa* genome in a pattern consistent with MITEs, being biased towards gene-rich regions and particularly in putative promoter regions.

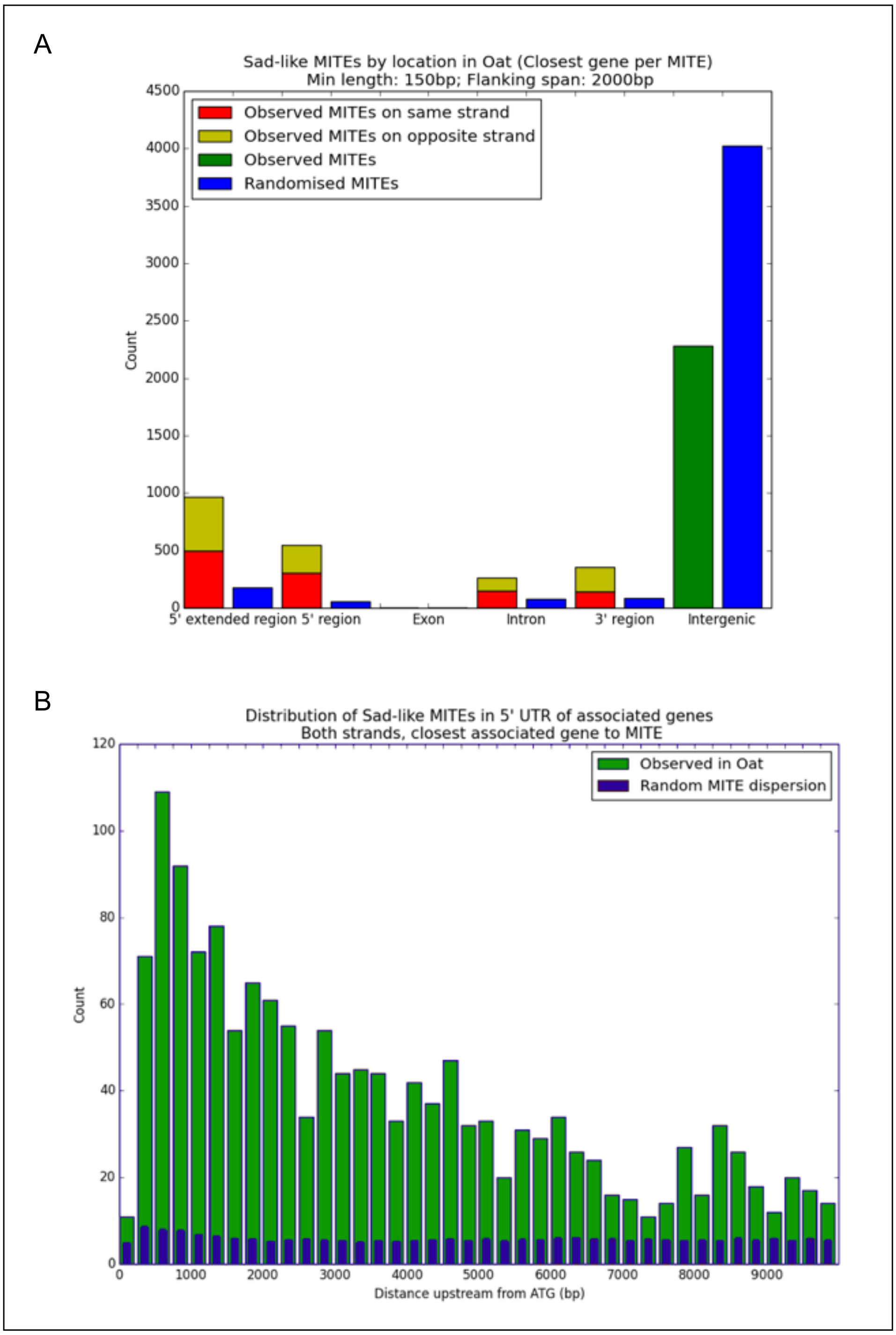


Figure 7.4 **Distribution of MITE-like sequences in the *A. strigosa* genome**

A) Distribution of elements throughout the genome with homology to the MITEs observed in the avenacin BGC. Where an element was found in proximity to a gene according to the definitions in 7.4 Methods, a classification was made as to whether the observed element was on the same or opposite strand as the gene. The observed distributions were compared to the data returned from hypothetical random distributions, shown in blue. B) Histogram of observed distribution of elements (in green) and randomised values (in blue) for those found 0-10kbp upstream of the start site of the co-located gene.

7.3.2 MITE-like sequences are not correlated with other BGCs or a conserved expression profile outside of the avenacin BGC

Genes with these elements present within the 2kbp upstream of the start site were assessed further to test if other examples of gene clustering or co-expression could be found. Outputs from plantiSMASH 1.0 analysis of the *A. strigosa* genome were cross-referenced with these genes. Whilst there were some cases of individual genes in putative BGCs that also had an element present in the promoter region, no BGC was found with more than one such gene other than the avenacin BGC (data not shown). To observe whether these MITE-like sequences were enriched in the promoter regions of genes that shared a similar functional role to those in the avenacin BGC, gene ontology (GO) term enrichment was also carried out on all genes with these elements in the 2kbp upstream of the gene start site. No particular conserved roles were found consistent with triterpene biosynthetic pathways, though genes involved in DNA binding and transcription factor activity did appear to be enriched (Table A2).

Co-expression analysis was carried out to observe if these elements were also present in other genes with a similar expression profile to the avenacin BGC. Expression levels of genes across six oat tissues were organised in a self-organising map (SOM) using the 'kohonen' package in R to group genes into bins of various expression profiles (Figure 7.5) [129]. The avenacin BGC expression profile was the only one with a particular enrichment of the MITE-like sequences, as shown by the clear enrichment in a single unit in Figure 7.5. However, it is only the avenacin BGC genes which contribute to this enrichment; no further genes with such elements within the putative promoter regions shared the same expression profile.

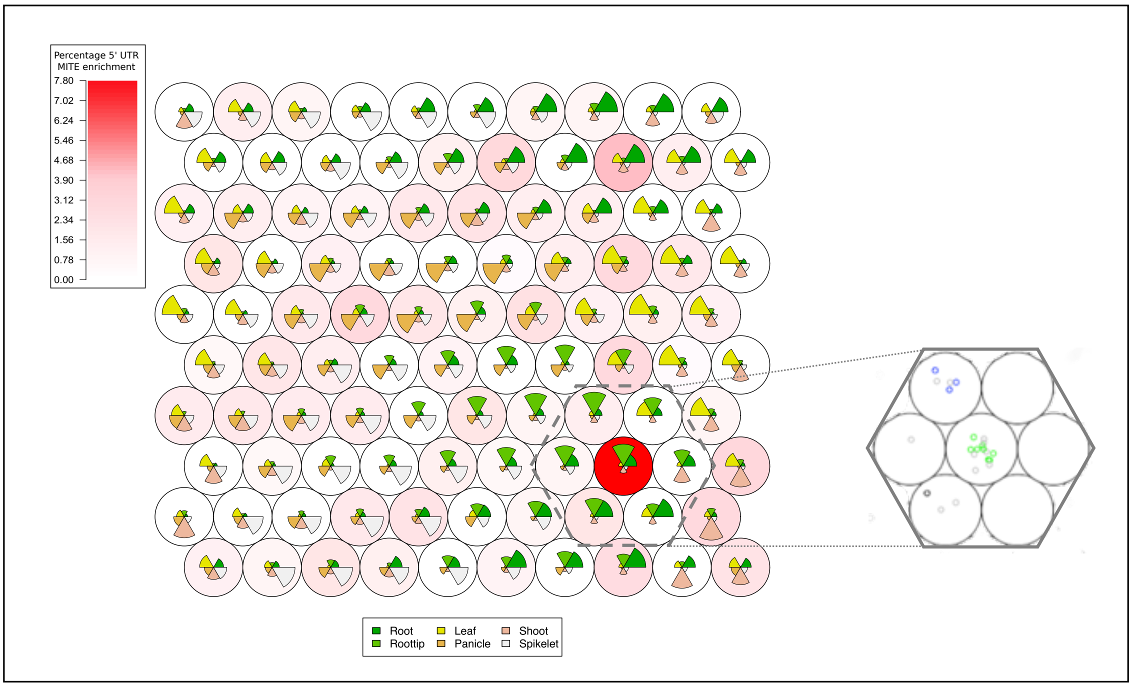


Figure 7.5 **SOM of gene expression profiles from *A. strigosa***
Colour of unit is determined by the proportion of genes assigned to that unit that contain MITE sequences within 0-2kbp upstream of the start codon (% enrichment as per key). Excerpt shows placement of genes in units. Coloured genes are the clustered avenacin biosynthetic genes. Blue: C30 P450, *Sad2*, *AAT*. Green: *Sad1*, *Sad7*, *Sad10*, *Sad9*, C21 P450, C30 P450, *UGT74H7*, *AsUGT91*, *AsTG*. Black: C23 P450

7.4 Conclusions

Despite their conservation of sequence and relative position to five genes in the avenacin BGC, the data here demonstrates that it is still unclear as to the role, if any, they may have or have had. The potential for the role of MITEs in BGC regulation and/or formation is attractive and has been demonstrated by their significant enrichment at loci with *OSC-CYP* gene pairs [37]. Future work may therefore broaden scope of TE analysis across numerous plant species to investigate whether there are any further examples of 'guilt by association'. It is hoped that any common features between examples may highlight potential mechanisms and allow hypotheses to be developed.

Given the rapid evolutionary nature of TEs and the scale of their distribution amongst plant genomes, great care must be taken in any studies such as these until quantifiable evidence can be found of their roles [126,128]. Nonetheless, any indication of similar 'flags' for generalised mechanisms of BGC formation and/or regulation in plant genomes merits continual investigation, as the presence of such phenomenon would open the door for truly unbiased and global genome mining for BGCs and provide a fascinating insight into the control of plant genome organisation.

Chapter 8. General Discussion

In this thesis, extant tools for BGC mining were reviewed and plantiSMASH 1.0 [34] was used to analyse a range of plant genome data in order to investigate the quality of results currently available for putative triterpene BGCs. A number of limitations were identified, which were suspected to be able to be overcome given current knowledge of key enzyme families (Chapter 2). Specifically, more refined levels of classification and subsequent product prediction were thought to be likely available.

An in-depth analysis of OSCs was then carried out (Chapter 3), which catalyse the first committed step for triterpene biosynthesis and open the way for the production of vast diversity of triterpenoid compounds found in nature [41]. OSCs control a nuanced chemical reaction and the sequence-function relationship is not well understood [66,67,69,133]. Through the development of tools to increase utility of unannotated genome data and subsequent phylogenetic study, it was observed that plant clades contain characteristic repertoires of OSC groups. Some of these represent convergent evolution of OSC function and others are indicative of key gene family radiations to access specialised chemistry which is unique to specific, but generally broad, plant clades. Through inspection of these data and the creation of a profile-based classification tools, it was concluded that phylogeny can be used to predict OSC function and help identify likely candidates of interest for further investigation.

Building from the previously studied co-evolutionary relationships between TPSs and CYPs, a large-scale analysis was carried out to observe patterns of OSC-CYP gene pair co-location across the Viridiplantae (Chapter 4). To achieve this, a rapid and accurate tool was developed to classify CYPs without having to build and inspect phylogenies. It was found that OSC-CYP co-location is highly diverse across plant clades, but does reconstitute known BGCs and functional relationships observed in given species. The data suggested that previous conclusions regarding fundamental differences between monocot and dicot TPS-CYP co-evolutionary relationships [36] were likely due to low sample size and subsequent sampling bias.

Data from numerous experiments involving GT1 function which were recently collated into a comprehensive review [30] were incorporated into a predictive tool to allow rapid analyses of candidate genes (Chapter 5). This was validated against putative enzyme sequences from *Quillaja saponaria*, a species which makes numerous saponins including QS-21 – a vaccine adjuvant with numerous non-glucose sugar moieties [108,112]. Furthermore, inspection of gene clustering for the QS-21 pathway revealed that GT1s were commonly co-localised with other biosynthetic genes, but the earlier pathway steps were not. This has wider implications for the creation and maintenance of BGCs in plants, especially those that produce a broad range of a given family of specialised metabolites in comparison to pathways for a very specific or atypical compound.

After these detailed studies of known triterpene biosynthetic gene families, an investigation was made into broadening the scope of plant BGC mining (Chapter 6). The limits of *ab initio* gene prediction tools combined with gene-density BGC mining parameters were found, though distance-based approaches were preliminarily successful. Unbiased studies of OSC neighbourhoods demonstrated the complexity and variability of gene co-location across plant species, but highlighted numerous new avenues for exploration. Firstly, glycosyl-hydrolases, a subfamily of which has been newly characterised to act as trans-glucosidases in triterpene biosynthetic pathways [115], were re-discovered via this approach. Furthermore, a range of putative regulatory and transporter gene families were identified, none of which have so far been identified as components of plant BGCs.

Finally, a focussed genome-wide analysis of *A. strigosa* was made in order to assess the prevalence of MITE-like sequences that were implicated in the regulation and/or assembly of the avenacin BGC (Chapter 7), as well as BGCs across a range of plant species [37]. It was postulated that such elements might serve as generalised signals for BGC formation, a discovery that would instantly allow new approaches to be made towards BGC mining as well as directly provide a mechanistic means for BGC creation. However, it was found that these elements are not correlated with other BGCs, did not confer a conserved expression pattern outside of the avenacin BGC genes nor do they exhibit a conserved sequence to other homologous elements in the genome.

The outcomes of this thesis can therefore be generally grouped into two areas of fundamental scientific interest. The first is the understanding of the evolution and dynamics of plant triterpene BGCs and their constituent genes via broad-scale genome mining across the Viridiplantae. The second is in the development and implementation of classification and predictive tools for putative biosynthetic enzymes in the context of a synthetic biology approach to metabolic engineering.

Classification and functional prediction tools were developed for three critically important enzyme families for triterpene biosynthetic pathways. The requirement for such tools to be able to form part of systematic, high-throughput pipelines was a priority throughout the work described here, and it is hoped that they may form part of future comprehensive BGC mining approaches. Whilst this thesis has necessarily focussed on the analysis of genome data, all of the tools described here may be applied to transcriptome data, where the wider taxonomic range is conducive to the discovery of particularly novel and diverse candidate enzymes [25]. Used in conjunction with gene synthesis, rapid enzyme characterisation platforms [21] and subsequent adjustment of the predictive models used for candidate selection, these advances represent an attempt to build a fundamentally important section of a genuine synthetic biology approach to plant triterpene metabolic engineering [26]. The ultimate aim in this context is the ability to produce target molecules 'on-demand', though how realistic such a goal this is remains to be seen.

Using triterpene biosynthetic enzymes as exemplars for the study of clustering dynamics and variety across the Viridiplantae has proven to be successful. It has been observed that plant clades often have signature patterns of both biosynthetic gene sub-families and patterns of co-location patterns of specific enzymes, all of which has been consistent with known, characterised

BGCs and pathways. Nonetheless, as is so often the case in nature, there appear to be few ‘hard rules’ when it comes to BGCs. The data presented here and in the recent literature suggests that the recruitment of genes as part of a BGC is far more dynamic and mutable than perhaps previously thought [35]. In addition to this, the areas between non-clustered biosynthetic pathways and totally clustered ones are only beginning to be explored. It will be particularly fascinating to learn how such patterns and distributions of gene families impact the biosynthetic potential of a given species, and how tight spatio-temporal regulation of specialised metabolite production is maintained across this spectrum of fluctuating gene organisation.

The volume of sequence data that will be available for analysis in the immediate future is astounding [24]. How we effectively handle genomes from tens of thousands of plant species and millions of transcriptomic datasets is a challenge we must solve now in order to make the most of the data available to us. In this manner, progress can only be made with continual efforts to build and iteratively improve systematic computational approaches and, critically, ensure that they are grounded in the reality of the lab. It is hoped that this thesis forms a small part of such a process.

Chapter 9. General Methods

9.1 General code and software

Analyses were carried out using Python [134] and R [131]. Specific modules used for handling, analysis and presentation of biological data were BioPython [135] (including the following packages: ETE toolkit, matplotlib, seaborn, Beautiful Soup, numpy and scipy), TopGO [130] and kohonen [129], Similarly, software that was used includes HMMer [62], BLAST [136], Selenoprofiles [74], Augustus [70], Exonerate [73], GeneWise [137], GlimmerHMM [71], genBlastG [72], CD-HIT [96], MAFFT [75], FastTree [138], RaXML [76], MrBayes [77], Dendroscope [139], seaview [140] and MEGA [141], all of which were installed according to the developers' instructions and run with default parameters unless otherwise stated.

9.2 Alignments and phylogentics

For alignments of conserved gene families, MAFFT [75] was used using the global pairwise alignment model, or otherwise default parameters for more diverse sequences. Trees were generated using FastTree [138], RaXML [76] and MrBayes [77], using default parameters unless otherwise stated.

Bibliography

1. Huang AC, Jiang T, Liu YX, Bai YC, Reed J, Qu B, Goossens A, Nützmann HW, Bai Y, Osbourn A: **A specialized metabolic network selectively modulates Arabidopsis root microbiota**. *Science (80-)* 2019, **364**:eaau6389.
2. Wang S, Alseekh S, Fernie AR, Luo J: **The Structure and Function of Major Plant Metabolite Modifications**. *Mol Plant* 2019, **12**:899–919.
3. Knudsen C, Gallage NJ, Hansen CC, Møller BL, Laursen T: **Dynamic metabolic solutions to the sessile life style of plants**. *Nat Prod Rep* 2018, **35**:1140–1155.
4. Panche AN, Diwan AD, Chandra SR: **Flavonoids: An overview**. *J Nutr Sci* 2016, **5**:e47.
5. Faizal A, Geelen D: **Saponins and their role in biological processes in plants**. *Phytochem Rev* 2013, **12**:877–893.
6. Singh B, Sharma RA: **Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications**. *3 Biotech* 2015, **5**:129–151.
7. Robles M, Aregullin M, West J, Rodriguez E: **Recent Studies on the Zoopharmacognosy, Pharmacology and Neurotoxicology of Sesquiterpene Lactones***. *Planta Med* 1995, **61**:199–203.
8. Mason PA, Bernardo MA, Singer MS: **A mixed diet of toxic plants enables increased feeding and anti-predator defense by an insect herbivore**. *Oecologia* 2014, **176**:477–486.
9. Singer MS, Mace KC, Bernays EA: **Self-medication as adaptive plasticity: Increased ingestion of plant toxins by parasitized caterpillars**. *PLoS One* 2009, **4**:e4796.
10. Aboelsoud NH: **Herbal medicine in ancient Egypt**. *J Med Plants Res* 2010, **4**:082–086.
11. Hill DJ: **Is there a future for natural dyes?** *Rev Prog Color Relat Top* 1997, **27**:18–25.
12. Osbourn A: **Saponins and plant defence — a soap story**. *Trends Plant Sci* 1996, **1**:4–9.
13. Van Setten DC, Van De Werken G: **Molecular structures of saponins from Quillaja saponaria molina**. In *Advances in Experimental Medicine and Biology*. . Springer, Boston, MA; 1996:185–193.
14. Afab K, Shaheen F, Mohammad FV, Noorwala M, Ahmad VU: **Saponins Used in Traditional and Modern Medicine**. *Advences Exp Med Biol* 1996, **404**:429–442.
15. Jia Z, Koike K, Nikaido T: **Major triterpenoid saponins from Saponaria officinalis**. *J Nat Prod* 1998, **61**:1368–1373.
16. Zhou Y, Ma Y, Zeng J, Duan L, Xue X, Wang H, Lin T, Liu Z, Zeng K, Zhong Y, et al.: **Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae**. *Nat Plants* 2016, **2**:16183.
17. Rosenthal JP, Dirzo R: **Effects of life history, domestication and agronomic selection on plant defence against insects: Evidence from maizes and wild relatives**. *Evol Ecol* 1997, **11**:337–355.

18. Whitehead SR, Turcotte MM, Poveda K: **Domestication impacts on plant-herbivore interactions: A meta-analysis**. *Philos Trans R Soc B Biol Sci* 2017, **372**:20160034.

19. Sivapalan T, Melchini A, Saha S, Needs PW, Traka MH, Tapp H, Dainty JR, Mithen RF: **Bioavailability of Glucoraphanin and Sulforaphane from High-Glucoraphanin Broccoli**. *Mol Nutr Food Res* 2018, **62**:1700911.

20. Butelli E, Titta L, Giorgio M, Mock HP, Matros A, Peterek S, Schijlen EGWM, Hall RD, Bovy AG, Luo J, et al.: **Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors**. *Nat Biotechnol* 2008, **26**:1301–1308.

21. Reed J, Stephenson MJ, Miettinen K, Brouwer B, Leveau A, Brett P, Goss RJM, Goossens A, O'Connell MA, Osbourn A, et al.: **A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules**. *Metab Eng* 2017, **42**:185–193.

22. Dai Z, Liu Y, Guo J, Huang L, Zhang X: **Yeast synthetic biology for high-value metabolites**. *FEMS Yeast Res* 2015, **15**:1–11.

23. Bolger M, Schwacke R, Gundlach H, Schmutzer T, Chen J, Arend D, Oppermann M, Weise S, Lange M, Fiorani F, et al.: **From plant genomes to phenotypes**. *J Biotechnol* 2017, **261**:46–52.

24. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev E V, Sun W, et al.: **10KP: A phylodiverse genome sequencing plan.** *Gigascience* 2018, **7**:1–9.

25. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al.: **One thousand plant transcriptomes and the phylogenomics of green plants**. *Nature* 2019, **574**:679–685.

26. Owen C, Patron NJ, Huang A, Osbourn A: **Harnessing plant metabolic diversity**. *Curr Opin Chem Biol* 2017, **40**:24–30.

27. Tissier A, Ziegler J, Vogt T: **Specialized Plant Metabolites: Diversity and Biosynthesis**. In *Ecological Biochemistry: Environmental and Interspecies Interactions*. . Wiley-VCH Verlag; 2015:14–37.

28. Pott DM, Osorio S, Vallarino JG: **From central to specialized metabolism: An overview of some secondary compounds derived from the primary metabolism for their role in conferring nutritional and organoleptic characteristics to fruit**. *Front Plant Sci* 2019, **10**:835.

29. Nelson D, Werck-Reichhart D: **A P450-centric view of plant evolution**. *Plant J* 2011, **66**:194–211.

30. Louveau T, Osbourn A: **The sweet side of plant-specialized metabolism**. *Cold Spring Harb Perspect Biol* 2019, **11**:a034744.

31. Kooke R, Keurentjes JJB: **Multi-dimensional regulation of metabolic networks**

shaping plant development and performance. *J Exp Bot* 2012, **63**:3353–65.

32. Nützmann HW, Huang A, Osbourn A: **Plant metabolic clusters – from genetics to genomics**. *New Phytol* 2016, **211**:771–789.

33. Nützmann H-W, Scazzocchio C, Osbourn A: **Metabolic Gene Clusters in Eukaryotes**. *Annu Rev Genet* 2018, **52**:159–183.

34. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH: **plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters.** *Nucleic Acids Res* 2017, **45**:W55–W63.

35. Liu Z, Suarez Duran HG, Harnvanichvech Y, Stephenson MJ, Schranz ME, Nelson D, Medema MH, Osbourn A: **Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae**. *New Phytol* 2019, **227**:1109–1123.

36. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, Osbourn A: **Investigation of terpene diversification across multiple sequenced plant genomes**. *Proc Natl Acad Sci* 2015, **112**:E81–E88.

37. Boutanaev AM, Osbourn AE: **Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots**. *Proc Natl Acad Sci U S A* 2018, **115**:E6650–E6658.

38. Kemen AC, Honkanen S, Melton RE, Findlay KC, Mugford ST, Hayashi K, Haralampidis K, Rosser SJ, Osbourn A: **Investigation of triterpene synthesis and regulation in oats reveals a role for β-amyrin in determining root epidermal cell patterning.** *Proc Natl Acad Sci U S A* 2014, **111**:8679–84.

39. Yu N, Nützmann HW, Macdonald JT, Moore B, Field B, Berriri S, Trick M, Rosser SJ, Kumar SV, Freemont PS, et al.: **Delineation of metabolic gene clusters in plant genomes by chromatin signatures**. *Nucleic Acids Res* 2016, **44**:2255–2265.

40. Chen F, Tholl D, Bohlmann J, Pichersky E: **The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom**. *Plant J* 2011, **66**:212–229.

41. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A: **Triterpene biosynthesis in plants.** *Annu Rev Plant Biol* 2014, **65**:225–57.

42. Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A: **A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants**. *Proc Natl Acad Sci U S A* 2004, **101**:8233–8238.

43. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, Osbourn A: **Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production.** *Proc Natl Acad Sci U S A* 2019, **116**:17096–17104.

44. Delis C, Krokida A, Georgiou S, Peña-Rodríguez LM, Kavroulakis N, Ioannou E,

Roussis V, Osbourn AE, Papadopoulou KK: **Role of lupeol synthase in Lotus japonicus nodule formation**. *New Phytol* 2011, **189**:335–346.

45. Go YS, Lee SB, Kim HJ, Kim J, Park H-Y, Kim J-K, Shibata K, Yokota T, Ohyama K, Muranaka T, et al.: **Identification of marneral synthase, which is critical for growth and development in Arabidopsis**. *Plant J* 2012, **72**:791–804.

46. Zhou S-F, Wang Y-Y, Zhe H, Yang Y, He Z: **Bardoxolone methyl (CDDO-Me) as a therapeutic agent: an update on its pharmacokinetic and pharmacodynamic properties**. *Drug Des Devel Ther* 2014, **8**:2075.

47. Ríos JL, Recio MC, Máñez S, Giner RM: **Natural triterpenoids as anti-inflammatory agents**. *Stud Nat Prod Chem* 2000, **22**:93–143.

48. Careaga VP, Bueno C, Muniain C, Alché L, Maier MS: **Antiproliferative, cytotoxic and hemolytic activities of a triterpene glycoside from Psolus patagonicus and its desulfated analog**. *Chemotherapy* 2009, **55**:60–68.

49. Zhang Y, Zhao L, Huang SW, Wang W, Song SJ: **Triterpene saponins with neuroprotective effects from the leaves of Diospyros kaki Thunb**. *Fitoterapia* 2018, **129**:138–144.

50. Yang L, Calingasan NY, Thomas B, Chaturvedi RK, Kiaei M, Wille EJ, Liby KT, Williams C, Royce D, Risingsong R, et al.: **Neuroprotective Effects of the Triterpenoid, CDDO Methyl Amide, a Potent Inducer of Nrf2-Mediated Transcription**. *PLoS One* 2009, **4**:e5757.

51. Yendo ACA, De Costa F, Gosmann G, Fett-Neto AG: **Production of plant bioactive Triterpenoid saponins: Elicitation strategies and target genes to improve yields**. *Mol Biotechnol* 2010, **46**:94–104.

52. Fiore C, Eisenhut M, Krausse R, Ragazzi E, Pellati D, Armanini D, Bielenberg J: **Antiviral effects of Glycyrrhiza species**. *Phyther Res* 2008, **22**:141–148.

53. Zhu D, Tuo W: **QS-21: A Potent Vaccine Adjuvant.** *Nat Prod Chem Res* 2016, **3**.

54. Bhatt JP: **Neurodepressive action of a piscicidal glycoside of plant, Aesculus indica (Colebr.) in fish.** *Indian J Exp Biol* 1992, **30**:437–9.

55. Sainsbury F, Lomonossoff GP: **Transient expressions of synthetic biology in plants**. *Curr Opin Plant Biol* 2014, **19**:1–7.

56. Medema MH, Osbourn A: **Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways**. *Nat Prod Rep* 2016, **492**:138–142.

57. Witjes L, Kooke R, Van Der Hooft JJJ, De Vos RCH, Keurentjes JJB, Medema MH, Nijveen H: **A genetical metabolomics approach for bioprospecting plant biosynthetic gene clusters**. *BMC Res Notes* 2019, **12**:1–5.

58. Töpfer N, Fuchs L-M, Aharoni A: **The PhytoClust tool for metabolic gene clusters discovery in plant genomes**. *Nucleic Acids Res* 2017, **45**:7049–7063.

59. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, et al.: **Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants.** *Plant Physiol* 2017, **173**:2041–2059.

60. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R: **antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences**. *Nucleic Acids Res* 2011, **39**:W339–W346.

61. Chae L, Kim T, Nilo-Poyanco R, Rhee SY: **Genomic Signatures of Specialized Metabolism in Plants**. *Science (80-)* 2014, **344**:510–513.

62. Eddy SR: **Accelerated profile HMM searches**. *PLoS Comput Biol* 2011, **7**:e1002195.

63. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al.: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**:D222–D230.

64. Leitch IJ, Leitch AR: **Genome size diversity and evolution in land plants**. In *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Edited by Greilhuber J, Dolezel J, Wendel JF. Springer-Verlag; 2013:307–322.

65. Medema MH, Fischbach MA: **Computational approaches to natural product discovery.** *Nat Chem Biol* 2015, **11**:639–48.

66. Xue Z, Duan L, Liu D, Guo J, Ge S, Dicks J, Ómáille P, Osbourn A, Qi X: **Divergent evolution of oxidosqualene cyclases in plants**. *New Phytol* 2012, **193**:1022–1038.

67. Xue Z, Tan Z, Huang A, Zhou Y, Sun J, Wang X, Thimmappa RB, Stephenson MJ, Osbourn A, Qi X: **Identification of key amino acid residues determining product specificity of 2,3-oxidosqualene cyclase in *Oryza* species**. *New Phytol* 2018, **218**:1076–1088.

68. Tian BX, Wallrapp FH, Holiday GL, Chow JY, Babbitt PC, Poulter CD, Jacobson MP: **Predicting the Functions and Specificity of Triterpenoid Synthases: A Mechanism-Based Multi-intermediate Docking Approach**. *PLoS Comput Biol* 2014, **10**:e1003874.

69. Salmon M, Thimmappa RB, Minto RE, Melton RE, Hughes RK, O'Maille PE, Hemmings AM, Osbourn A: **A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases.** *Proc Natl Acad Sci U S A* 2016, **113**:E4407-14.

70. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: A web server for gene finding in eukaryotes**. *Nucleic Acids Res* 2004, **32**:W309–W312.

71. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**:2878–2879.

72. She R, Chu JSC, Uyar B, Wang J, Wang K, Chen N: **genBlastG: Using BLAST searches to build homologous gene models**. *Bioinformatics* 2011, **27**:2141–2143.

73. Slater GSC, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6**:31.

74. Mariotti M, Guigó R: **Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes**. *Bioinformatics* 2010, **26**:2656–2663.

75. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: Improvements in performance and usability**. *Mol Biol Evol* 2013, **30**:772–780.

76. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30**:1312–3.

77. Huelsenbeck JP, Ronquist F: **Bayesian Analysis of Molecular Evolution Using MrBayes**. In *Statistical Methods in Molecular Evolution.* . Springer-Verlag; 2005:183–226.

78. Araki T, Saga Y, Marugami M, Otaka J, Araya H, Saito K, Yamazaki M, Suzuki H, Kushiro T: **Onocerin Biosynthesis Requires Two Highly Dedicated Triterpene Cyclases in a Fern Lycopodium clavatum**. *ChemBioChem* 2016, **17**:288–290.

79. Jian CC, Ming HC, Rui LN, Cordel GA, Qiuz SX: **Cucurbitacins and cucurbitane glycosides: Structures and biological activities**. *Nat Prod Rep* 2005, **22**:386–399.

80. Hoshino T: **β-Amyrin biosynthesis: catalytic mechanism and substrate recognition**. *Org Biomol Chem* 2017, **15**:2869–2891.

81. Xue Z, Xu X, Zhou Y, Wang X, Zhang Y, Liu D, Zhao B, Duan L, Qi X: **Deficiency of a triterpene pathway results in humidity-sensitive genic male sterility in rice**. *Nat Commun* 2018, **9**:604.

82. Xu Y, Zhang Z, Wang M, Wei J, Chen H, Gao Z, Sui C, Luo H, Zhang X, Yang Y, et al.: **Identification of genes related to agarwood formation: Transcriptome analysis of healthy and wounded tissues of Aquilaria sinensis**. *BMC Genomics* 2013, **14**:227.

83. Liu Y, Chen H, Yang Y, Zhang Z, Wei J, Meng H, Chen W, Feng J, Gan B, Chen X, et al.: **Whole-tree agarwood-inducing technique: An efficient novel technique for producing high-quality agarwood in cultivated Aquilaria sinensis trees**. *Molecules* 2013, **18**:3086–3106.

84. Chen HQ, Wei JH, Yang JS, Zhang Z, Yang Y, Gao ZH, Sui C, Gong B: **Chemical constituents of agarwood originating from the endemic genus Aquilaria plants**. *Chem Biodivers* 2012, **9**:236–250.

85. Chen CH, Kuo TCY, Yang MH, Chien TY, Chu MJ, Huang LC, Chen CY, Lo HF, Jeng ST, Chen LFO: **Identification of cucurbitacins and assembly of a draft genome for Aquilaria agallocha**. *BMC Genomics* 2014, **15**:1–11.

86. Christiansen JA: **Note on Negative Catalysis**. *J Phys Chem* 1924, **28**:145–148.

87. Rétey J: **Enzymic Reaction Selectivity by Negative Catalysis or How Do Enzymes Deal with Highly Reactive Intermediates?** *Angew Chemie Int Ed English* 1990, **29**:355–361.

88. Breslow R: **How Do Imidazole Groups Catalyze the Cleavage of RNA in Enzyme Models and in Enzymes? Evidence from “Negative Catalysis.”** *Acc Chem Res* 1991, **24**:317–324.

89. Frickey T, Kannenberg E: **Phylogenetic analysis of the triterpene cyclase protein family in prokaryotes and eukaryotes suggests bidirectional lateral gene transfer**. *Environ Microbiol* 2009, **11**:1224–1241.

90. Goodman J: **Computer Software Review: Reaxys.** *J Chem Inf Model* 2009, **49**:2897–2898.

91. Ridley DD: *Information Retrieval: SciFinder® Second Edition*. John Wiley & Sons Ltd; 2009.

92. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I: **CATH: An expanded resource to predict protein function through structure and sequence**. *Nucleic Acids Res* 2017, **45**:D289–D295.

93. Nelson DR: **The cytochrome p450 homepage.** *Hum Genomics* 2009, **4**:59–65.

94. Ghosh S: **Triterpene structural diversification by plant cytochrome P450 enzymes**. *Front Plant Sci* 2017, **8**:1886.

95. Buchholz PCF, Vogel C, Reusch W, Pohl M, Rother D, Spieß AC, Pleiss J: **BioCatNet: A Database System for the Integration of Enzyme Sequences and Biocatalytic Experiments**. *ChemBioChem* 2016, **17**:2093–2098.

96. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data.** *Bioinformatics* 2012, **28**:3150–2.

97. Sun G, Xu Y, Liu H, Sun T, Zhang J, Hettenhausen C, Shen G, Qi J, Qin Y, Li J, et al.: **Large-scale gene losses underlie the genome evolution of parasitic plant Cuscuta australis**. *Nat Commun* 2018, **9**:1–8.

98. Suzuki M, Xiang T, Ohyama K, Seki H, Saito K, Muranaka T, Hayashi H, Katsube Y, Kushiro T, Shibuya M, et al.: **Lanosterol Synthase in Dicotyledonous Plants**. *Plant Cell Physiol* 2006, **47**:565–571.

99. Shang Y, Ma Y, Zhou Y, Zhang H, Duan L, Chen H, Zeng J, Zhou Q, Wang S, Gu W, et al.: **Biosynthesis, regulation, and domestication of bitterness in cucumber**. *Science (80-)* 2014, **346**:1084–1088.

100. Rupasinghe S, Schuler MA: **Homology modeling of plant cytochrome P450s**. *Phytochem Rev* 2006, **5**:473–505.

101. Su B-H, Tu Y, Lin C, Shao C-Y, Lin OA, Tseng YJ: **Rule-Based Prediction Models of Cytochrome P450 Inhibition**. *J Chem Inf Model* 2015, **55**:1426–1434.

102. Li Q, Fang Y, Li X, Zhang H, Liu M, Yang H, Kang Z, Li Y, Wang Y: **Mechanism of the plant cytochrome P450 for herbicide resistance: a modelling study**. *J Enzyme Inhib Med Chem* 2013, **28**:1182–1191.

103. Miettinen K, Pollier J, Buyst D, Arendt P, Csuk R, Sommerwerk S, Moses T, Mertens J,

Sonawane PD, Pauwels L, et al.: **The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis**. *Nat Commun* 2017, **8**:1–13.

104. Seki H, Sawai S, Ohyama K, Mizutani M, Ohnishi T, Sudo H, Fukushima EO, Akashi T, Aoki T, Saito K, et al.: **Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin**. *Plant Cell* 2011, **23**:4112–4123.

105. Sawai S, Saito K: **Triterpenoid biosynthesis and engineering in plants.** *Front Plant Sci* 2011, **2**:25.

106. Ross J, Yi L, Eng-Kiat L, Bowles D: **Protein family review Higher plant glycosyltransferases**. *Genome Biol* 2001, **2**:300–4.

107. Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S: **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**. *Plant J* 2012, **69**:1030–1042.

108. Fleck JD, Betti AH, Da Silva FP, Troian EA, Olivaro C, Ferreira F, Verza SG: **Saponins from Quillaja saponaria and Quillaja brasiliensis: Particular Chemical Characteristics and Biological Activities**. *Molecules* 2019, **24**:171.

109. Copaja S V., Blackburn C, Carmona R: **Variation of saponin contents in Quillaja saponica molina**. *Wood Sci Technol* 2003, **37**:103–108.

110. Basu N, Rastogi RP: **Triterpenoid saponins and sapogenins**. *Phytochemistry* 1967, **6**:1249–1270.

111. Liu G, Anderson C, Scaltreto H, Barbon J, Kensil CR: **QS-21 structure/function studies: Effect of acylation on adjuvant activity**. *Vaccine* 2002, **20**:2808–2815.

112. Read Kensil C, Kammer R: **QS-21: A water-soluble triterpene glycoside adjuvant**. *Expert Opin Investig Drugs* 1998, **7**:1475–1482.

113. Yang M, Fehl C, Lees K V., Lim EK, Offen WA, Davies GJ, Bowles DJ, Davidson MG, Roberts SJ, Davis BG: **Functional and informatics analysis enables glycosyltransferase activity prediction**. *Nat Chem Biol* 2018, **14**:1109–1117.

114. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, et al.: **Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters**. *Cell* 2014, **158**:412–421.

115. Orme A, Louveau T, Stephenson MJ, Appelhagen I, Melton R, Cheema J, Li Y, Zhao Q, Zhang L, Fan D, et al.: **A noncanonical vacuolar sugar transferase required for biosynthesis of antimicrobial defense compounds in oat**. *Proc Natl Acad Sci U S A* 2019, **116**:27105–27114.

116. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28**:45–8.

117. Huang AC, Kautsar SA, Hong YJ, Medema MH, Bond AD, Tantillo DJ, Osbourn A: **Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae through genome mining reveals convergent evolution**. *Proc Natl Acad Sci* 2017, **114**:E6005–E6014.

118. Rudolf JD, Chang C-Y: **Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases**. *Nat Prod Rep* 2020, **37**:425–463.

119. Somerville C: **Cellulose Synthesis in Higher Plants**. *Annu Rev Cell Dev Biol* 2006, **22**:53–78.

120. Frey M, Huber K, Park WJ, Sicker D, Lindberg P, Meeley RB, Simmons CR, Yalpani N, Gierl A: **A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-biosynthesis**. *Phytochemistry* 2003, **62**:371–376.

121. Wang D, Wang H, Irfan M, Fan M, Lin F: **Structure and evolution analysis of pollen receptor-like kinase in Zea mays and Arabidopsis thaliana**. *Comput Biol Chem* 2014, **51**:63–70.

122. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.** *Nucleic Acids Res* 2009, **37**:D233-8.

123. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y: **dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.** *Nucleic Acids Res* 2018, **46**:W95–W101.

124. Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, Mukherjee S, Ancona N: **Comparative study of gene set enrichment methods**. *BMC Bioinformatics* 2009, **10**:275.

125. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R: **Avoiding the pitfalls of gene set enrichment analysis with SetRank**. *BMC Bioinformatics* 2017, **18**:151.

126. Feng Y: **Plant MITEs: useful tools for plant genetics and genomics.** *Genomics, proteomics Bioinforma / Beijing Genomics Inst* 2003, **1**:90–100.

127. Bennetzen JL: **Transposable elements, gene creation and genome rearrangement in flowering plants**. *Curr Opin Genet Dev* 2005, **15**:621–627.

128. Han Y, Wessler SR: **MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences.** *Nucleic Acids Res* 2010, **38**:e199.

129. Wehrens R: **Package "kohonen". R package**. In *R topics documented.* . CRAN; 2015.

130. Alexa A, Rahnenfuhrer J: **topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0**. In *R topics documented.* . CRAN; 2010.

131. R Development Core Team: *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. 2013.

132. Gouet P, Courcelle E, Stuart DI, Metoz F: **ESPript: analysis of multiple sequence**

alignments in PostScript. *Bioinformatics* 1999, **15**:305–308.

133. Xu R, Fazio GC, Matsuda SPT: **On the origins of triterpenoid skeletal diversity**. *Phytochemistry* 2004, **65**:261–291.

134. Van Rossum G, Drake Jr FL: *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

135. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al.: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**:1422–1423.

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

137. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**:988–995.

138. Price MN, Dehal PS, Arkin AP: **FastTree 2 - Approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**:e9490.

139. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: An interactive viewer for large phylogenetic trees**. *BMC Bioinformatics* 2007, **8**:1–6.

140. Gouy M, Guindon S, Gascuel O: **SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building**. *Mol Biol Evol* 2010, **27**:221–224.

141. Kumar S, Nei M, Dudley J, Tamura K: **MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences**. *Brief Bioinform* 2008, **9**:299–306.

Appendices

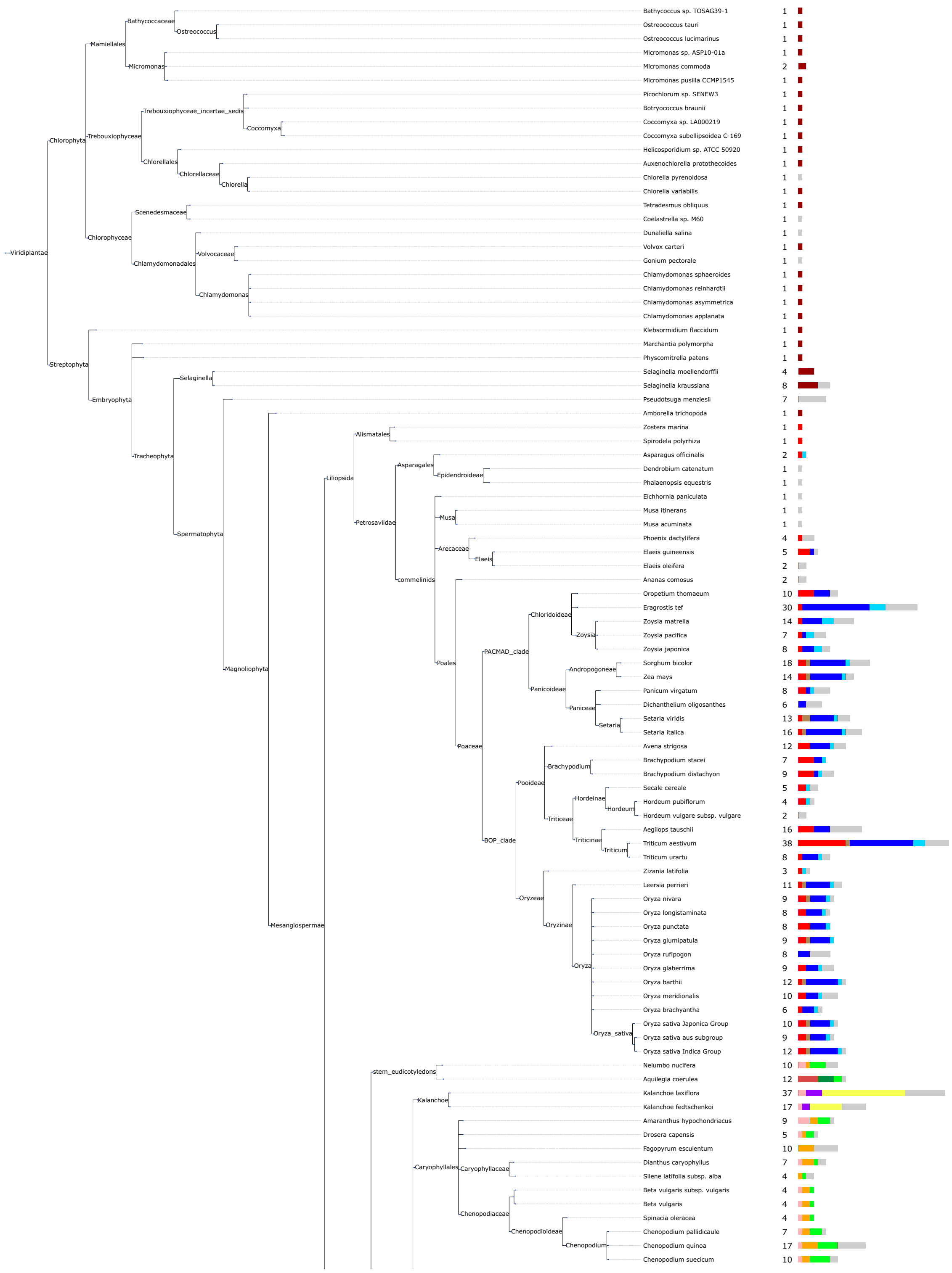


Figure A1

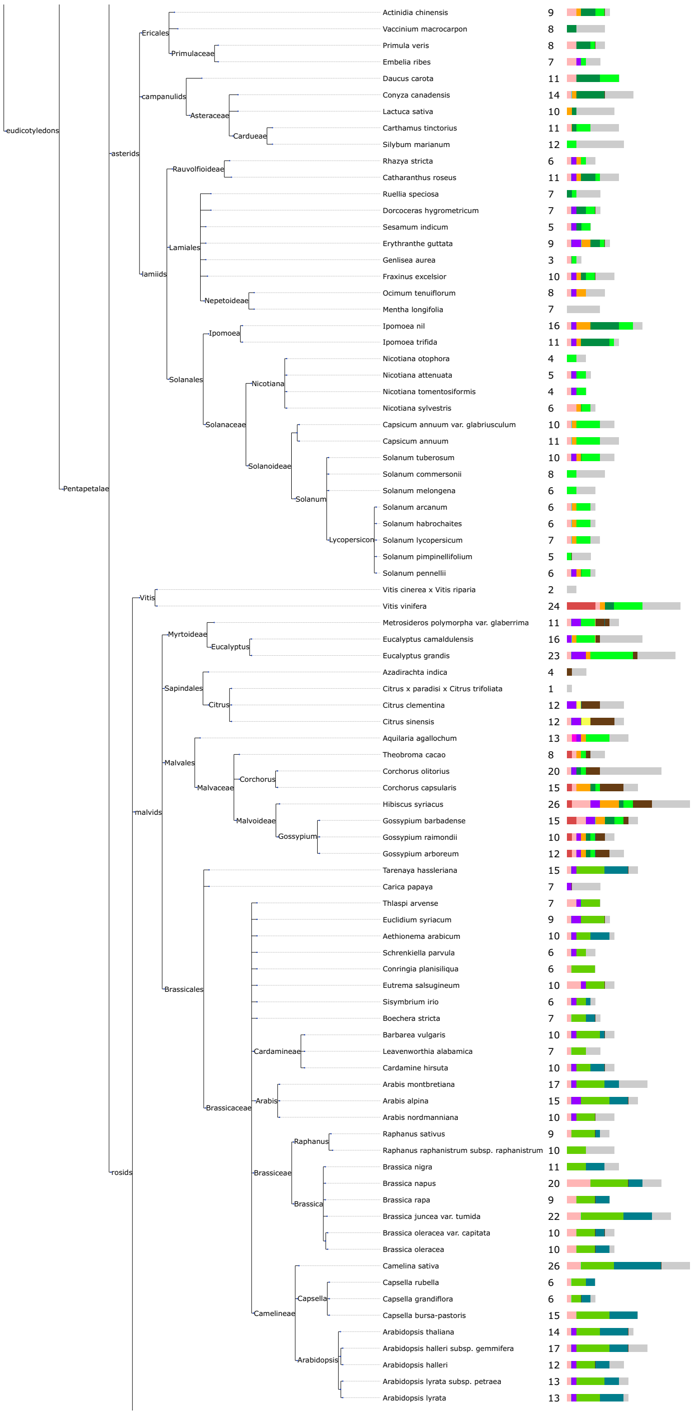


Figure A1 (cont.)

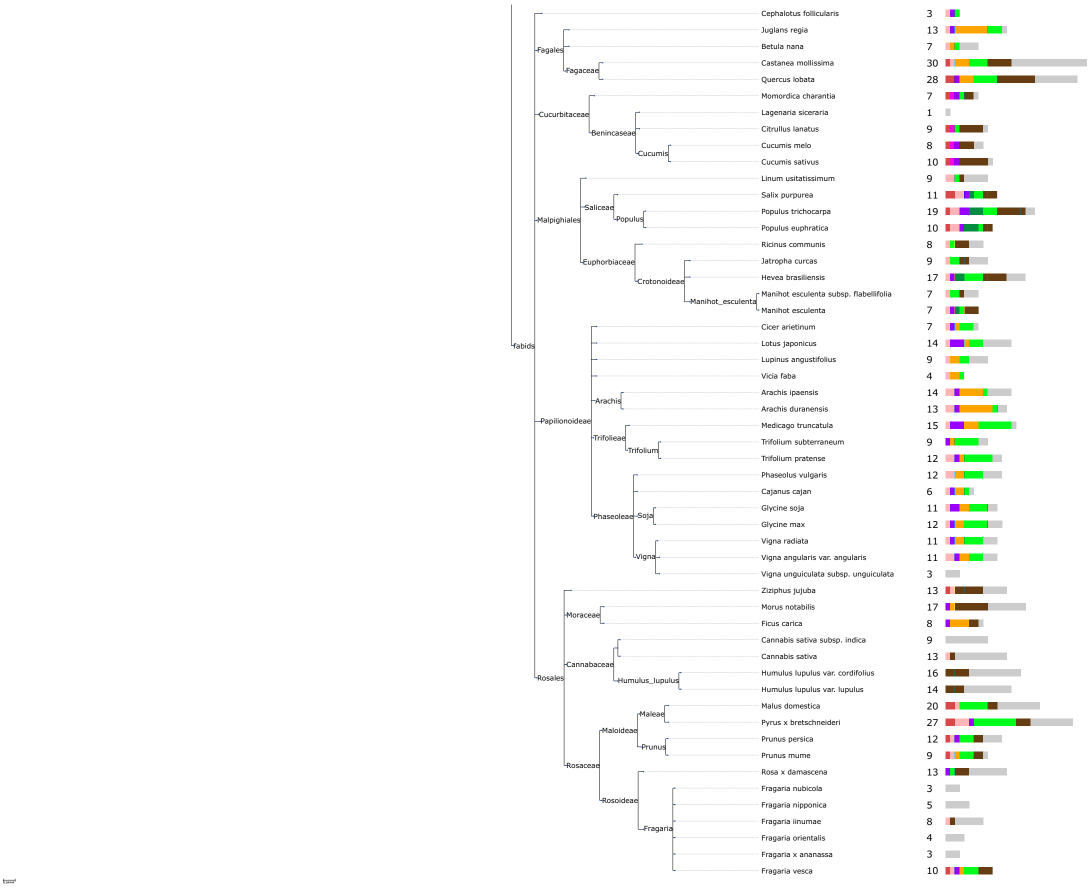


Figure A1 (cont.) **OSC 'fingerprinting' across the Viridiplantae**

Homology to conserved OSC groups can be used to predict the function of target candidates, discount candidates for desired functionalities and give snapshot as to the evolution and diversity of OSCs between species. Subtrees and key shown in Figure 3.7.

Table A1 **Viridiplantae genomes used in this thesis**

Species	*Genome ID*	*Number of contigs*	*Genome length (Mbp)*	*N50*	*Number of genes*	*Number of proteins*
Actinidia chinensis	GCA_000467755p1_Kiwifruit_v1	26721	604.2	58864	0	0
Actinidia chinensis	GCA_003024255p1_Red5_PS1_1p69p0	1234	553.8	18944233	33044	33115
Aegilops tauschii	GCA_000347335p1_ASM34733v1	429891	3313.7	68369	42871	33849
Aegilops tauschii	GCF_001957025p1_Aet_MR_1p0	68538	4327.3	468757	56362	55713
Aethionema arabicum	Aethionema_arabicum_formerly_known_as_Dick	3166	196.0	564741	22753	124430
Aethionema arabicum	GCA_000411095p1_VEGI_AA_v_1p0	18312	192.5	123806	0	0
Alnus glutinosa	GCA_003254965p1_ASM325496v1	167345	611.9	96611	0	0
Amaranthus hypochondriacus	Ahypochondriacus_315_v1p0	1777	361.4	396529	23038	23059
Amaranthus hypochondriacus	GCA_000753965p1_AHP_1p0	117340	502.1	42518	0	0
Amaranthus tuberculatus	GCA_000180655p1_ASM18065v1	15440	4.3	241	0	0
Amborella trichopoda	Atrichopoda_291_v1p0pgene_exons	5745	706.3	4927027	26846	109783
Amborella trichopoda	GCF_000471905p2_AMTR1p0	5746	706.5	4927027	19521	31494
Ananas comosus	Acomosus_321_v3	1322	361.2	12612916	27024	27024
Ananas comosus	GCA_001661175p1_ACMD2v1p0	8448	524.1	153084	23598	23598
Ananas comosus	GCF_001540865p1_ASM154086v1	3129	382.1	11759267	25758	35775
Apostasia shenzhenica	GCA_002786265p1_ASM278626v1	2985	348.7	3029156	21743	21743
Aquilaria agallochum	GCA_000696445p1_Aquilaria_agallocha_v1	27769	726.7	128399	0	0
Aquilegia coerulea	Acoerulea_322_v3p1	238	300.2	43571201	30023	43550
Aquilegia coerulea	GCA_002738505p1_Aquilegia_coerulea_v1	970	302.0	4232396	24823	41063
Arabidopsis halleri	Ahalleri_264_v1p1	6508	115.3	33068	25008	26911
Arabidopsis halleri subsp. gemmifera	GCA_900078215p1_Ahal2p2	2239	196.2	712249	0	0
Arabidopsis lyrata	Alyrata_384_v2p1pgene_exons	695	206.7	24464547	31073	169384
Arabidopsis lyrata	GCF_000004255p2_vp1p0	696	206.8	24464547	34365	39161
Arabidopsis lyrata subsp. petraea	GCA_000524985p1_Alyr_1p0	281536	203.0	7848	0	0
Arabidopsis thaliana	GCF_000001735p4_TAIR10p1	7	119.7	23459830	38093	48266
Arabis alpina	GCA_000733195p1_A_alpina_V4	27771	308.0	27950219	30690	23286

Arabis alpina	GCA_900128785p1_MPIPZpv5	8	311.6	36598175	0	0
Arabis montbretiana	GCA_001484125p1_ASM148412v1	28775	199.1	21621	0	0
Arabis nordmanniana	GCA_001484925p1_ASM148492v1	267228	342.3	4973	0	0
Arachis duranensis	GCA_001687015p1_ASM168701v1	20214	1076.0	149039	0	0
Arachis duranensis	GCF_000817695p2_Aradu1p1	1507	1084.3	110037037	45161	52826
Arachis hypogaea	GCA_003086295p1_arahypTifrunnerpgnm1pKYV3	20	2538.3	135150084	0	0
Arachis ipaensis	GCA_000816755p2_Araip1p1	548	1353.5	136175642	0	0
Arachis ipaensis	GCF_000816755p2_Araip1p1	548	1353.5	136175642	49814	57621
Arachis monticola	GCA_003063285p2_ASM306328v2	6909	2618.7	124915013	0	0
Argania spinosa	GCA_003260245p1_arg_spin_01	75327	670.1	49916	0	0
Artemisia annua	GCA_003112345p1_ASM311234v1	39400	1792.9	104891	63226	66918
Artocarpus camansi	GCA_002024485p1_Acamansi1p0	396025	631.3	2430	0	0
Asclepias syriaca	GCA_002018285p1_ASM201828v1	221855	236.8	1983	0	0
Asparagus officinalis	GCA_001876935p1_AspofpV1	11792	1187.5	131339754	27986	27395
Asparagus officinalis	GCF_001876935p1_AspofpV1	11792	1187.5	131339754	32237	36763
Atalantia buxifolia	GCA_002013935p1_ASM201393v1	25600	315.8	1073988	0	0
Auxenochlorella protothecoides	GCF_000733215p1_ASM73321v1	374	22.9	285543	7016	7014
Auxenochlorella pyrenoidosa	GCA_001430745p1_ASM143074v1	1346	57.0	1392758	0	0
Avena strigosa	OAT_v0p8	11080	3068.2	436998	51266	51266
Azadirachta indica	GCA_000439995p3_AzaInd2p1	126142	261.5	3491	0	0
Barbarea vulgaris	GCA_001920985p1_ASM192098v1	7810	167.4	56351	0	0
Bathycoccus prasinos	GCF_002220235p1_ASM222023v1	21	15.1	955652	7967	7900
Bathycoccus sp. TOSAG39-1	GCA_900128745p1_TOSAG39-1	2118	10.1	14082	0	0
Begonia fuchsioides	GCA_003255005p1_ASM325500v1	55006	373.9	154265	0	0
Berberis thunbergii	GCA_003290165p1_Bpthun_GenomeAssembly_v1	11815	2240.7	397058	0	0
Beta vulgaris	GCF_000511025p2_RefBeet-1p2p2	40246	566.6	34941034	28113	32874
Beta vulgaris subsp. vulgaris	GCA_000510975p1_RefBeet-1p1p1	43471	568.6	33895747	0	0
Betula nana	GCA_000327005p1_ASM32700v1	551915	564.0	18694	0	0
Betula pendula	GCA_900184695p1_Bpev01	5644	435.9	239696	0	0
Boechera stricta	Bstricta_278_v1p2	854	185.5	2333866	27416	29812
Boechera stricta	GCA_002079875p1_Bstricta_278_v1	1944	188.8	2187891	0	0
Boehmeria nivea	GCA_002937015p1_ASM293701v1	12775	344.6	1094501	0	0

Botryococcus braunii	GCA_002005505p1_B_braunii_Showa_v1	2751	184.4	372998	0	0
Brachypodium distachyon	Bdistachyon_314_v3p1	10	271.2	59130575	34310	52972
Brachypodium distachyon	GCA_000005505p4_Brachypodium_distachyon_v3p0	10	271.2	59130575	34310	52972
Brachypodium distachyon	GCF_000005505p3_Brachypodium_distachyon_v3p0	11	271.3	59130575	31335	37892
Brachypodium stacei	Bstacei_316_v1p1	27	233.8	23060899	29898	36357
Brassica cretica	GCA_003260655p1_B_cretica_A_v1	243461	412.5	2820	0	0
Brassica juncea var. tumida	GCA_001687265p1	9746	954.9	38841276	0	0
Brassica napus	Brassica_napuAST_PRJEB5043_v1p41pnonchromosomal	20899	848.2	680862	104573	101040
Brassica napus	GCF_000686985p1_Brassica_napus_assembly_v1p0	1376	930.5	41855496	112281	112890
Brassica napus	GCF_000686985p2_Bra_napus_v2p0	1471	976.2	45943547	119533	123467
Brassica nigra	GCA_001682895p1_ASM168289v1	2545	402.1	39062406	0	0
Brassica oleracea	GCF_000695525p1_BOL	32886	489.0	48366697	54054	56687
Brassica oleracea var. capitata	Boleraceacapitata_446_v1p0pgene_exons	9	385.0	40895475	35400	161719
Brassica oleracea var. capitata	GCA_000604025p1_BOL_v1p0	1816	514.4	1419759	0	0
Brassica rapa	BrapaFPsc_277_v1p3	669	298.6	28488603	40492	43370
Brassica rapa	GCA_000309985p2_ASM30998v2	70673	386.1	3377735	0	0
Brassica rapa	GCF_000309985p1_Brapa_1p0	40249	284.1	26286742	49056	51005
Cajanus cajan	GCA_000340665p1_Cpcajan_V1p0	36535	592.8	555764	50122	48331
Cajanus cajan	GCF_000340665p1_Cpcajan_V1p0	36536	593.0	555764	31841	38965
Calamus simplicifolius	GCA_900491605p1_Calamus_simplicifolius	5116	1960.8	803014	0	0
Camelina sativa	GCA_000496875p1_CamelinaSativa	15937	547.6	99217	0	0
Camelina sativa	GCA_000633955p1_Cs	37212	641.4	30099736	0	0
Camelina sativa	GCF_000633955p1_Cs	37212	641.4	30099736	96896	106267
Cannabis sativa	GCA_001865755p1_ASM186575v1	11110	585.8	128718	0	0
Cannabis sativa subsp. indica	GCA_001510005p1_ASM151000v1	311039	595.4	2649	0	0
Capsella bursa-pastoris	GCA_001974645p1_C_bursa_pastoris_nuclear	8186	268.4	627605	0	0
Capsella grandiflora	Cgrandiflora_266_v1p1	2710	100.4	122625	24805	26561
Capsella rubella	Crubella_183_v1p0pgene_exons	853	134.8	15060676	26521	148564
Capsella rubella	GCA_000375325p1_Caprub1_0	773	133.1	15040190	26776	28713
Capsella rubella	GCF_000375325p1_Caprub1_0	773	133.1	15040190	29301	34126

Capsicum annuum	GCA_000512255p1_PGAvp1p5	37989	3063.6	2472394	0	0
Capsicum annuum	GCA_000710875p1_Pepper_Zunla_1_Ref_v1p0	1625	2935.2	220335243	0	0
Capsicum annuum	GCF_000710875p1_Pepper_Zunla_1_Ref_v1p0	1627	2935.9	220335243	41729	45410
Capsicum annuum var. glabriusculum	GCA_000950795p1	3346	2768.1	200607515	0	0
Capsicum baccatum	GCA_002271885p2_ASM227188v2	23260	3215.6	229738584	35853	35853
Capsicum chinense	GCA_002271895p2_ASM227189v2	87978	3070.9	234238532	34974	34974
Cardamine hirsuta	C_hirsuta_v1	207	194.8	23393806	29458	37996
Carica papaya	Cpapaya_113_ASGPBv0p4	4111	324.5	1304607	27769	27793
Carica papaya	GCA_000150535p1_Papaya1p0	17764	369.8	1089885	0	0
Carica papaya	GCF_000150535p2_Papaya1p0	17766	370.4	1089885	20332	26103
Carnegiea gigantea	GCA_002740515p1_SGP5_Cgig_v1p3	57405	980.4	61549	0	0
Carthamus tinctorius	GCA_001633085p1_Safflower1	463906	661.9	3565	0	0
Castanea mollissima	GCA_000763605p1_ASM76360v1	133589	833.2	32186	0	0
Casuarina glauca	GCA_003255045p1_ASM325504v1	39787	282.8	912668	0	0
Catharanthus roseus	GCA_000949345p1_ASM94934v1	79302	522.7	26249	0	0
Cenchrus americanus	GCA_002174835p2_ASM217483v2	52033	1816.9	240570548	0	0
Cephalotus follicularis	GCA_001972305p1_Cfol_1p0	16307	1614.5	287498	36503	36667
Cercis canadensis	GCA_003255065p1_ASM325506v1	8828	329.3	419957	0	0
Chamaecrista fasciculata	GCA_003254925p1_ASM325492v1	56674	429.1	96643	0	0
Chenopodium pallidicaule	GCA_001687005p1_ASM168700v1	3013	337.0	356818	0	0
Chenopodium quinoa	GCA_001683475p1_ASM168347v1	3486	1333.4	3844283	0	0
Chenopodium quinoa	GCA_001742885p1_Cqu_r1p0	24845	1087.4	86941	0	0
Chenopodium quinoa	GCF_001683475p1_ASM168347v1	3487	1333.6	3844283	58882	63173
Chenopodium suecicum	GCA_001687025p1_ASM168702v1	11198	536.9	105389	0	0
Chlamydomonas applanata	GCA_001662365p1_Cap_assembly01	2533	78.5	105699	0	0
Chlamydomonas asymmetrica	GCA_001662385p1_Cas_assembly01	4102	141.9	114158	0	0
Chlamydomonas debaryana	GCA_001662405p1_Cde_assembly01	10139	120.4	27219	0	0
Chlamydomonas eustigma	GCA_002335675p1_Cpeustigma	520	66.6	465125	14112	14161
Chlamydomonas reinhardtii	Creinhardtii_281_v5p5	52	111.1	7783580	17741	19526
Chlamydomonas reinhardtii	GCA_000002595p3	53	111.1	7783580	17743	19528
Chlamydomonas reinhardtii	GCF_000002595p1_v3p0	1558	120.4	1695175	14488	14504
Chlamydomonas sphaeroides	GCA_001662425p1_Csp_assembly01	6890	122.2	44734	0	0
Chlorella sorokiniana	GCA_003130725p1_ASM313072v1	20	58.5	4091730	0	0

Chlorella sp. A99	GCA_003063905p1_ASM306390v1	82	40.9	1727419	0	0
Chlorella sp. ArM0029B	GCA_002896455p3_ArM29Bkp_1312	347	93.0	805067	0	0
Chlorella variabilis	GCF_000147415p1_v_1p0	414	46.2	1469606	9780	9780
Chlorella vulgaris	GCA_001021125p1_ASM102112v1	3600	37.3	27824	0	0
Chondrus crispus	GCA_000350225p2_ASM35022v2	926	105.0	242694	9843	9807
Chondrus crispus	GCF_000350225p1_ASM35022v2	926	105.0	242694	9843	9807
Cicer arietinum	GCA_000331145p1_ASM33114v1	7126	530.8	39989001	0	0
Cicer arietinum	GCA_000347275p2_ASM34727v2	38511	510.9	39901017	0	0
Cicer arietinum	GCF_000331145p1_ASM33114v1	7127	530.9	39989001	27889	33107
Cicer echinospermum	GCA_002896215p1_S2Drd065_v0p3	19348	644.7	206896	0	0
Cicer reticulatum	GCA_002896235p1_Besev079_v0p3	38802	715.4	109263	0	0
Cissus quadrangularis	GCA_002878655p1_ASM287865v1	125206	281.7	6999	0	0
Citrullus lanatus	GCA_000238415p1_CiLa_1p0	40248	321.0	26400	0	0
Citrus cavaleriei	GCA_002013975p2_ASM201397v2	14916	357.6	501435	0	0
Citrus clementina	Cclementina_182_v1p0pgene_exons	1398	301.4	31410901	24533	188707
Citrus clementina	GCA_000493195p1_Citrus_clementina_v1p0	1398	301.4	31410901	25000	34557
Citrus clementina	GCF_000493195p1_Citrus_clementina_v1p0	1398	301.4	31410901	27326	32586
Citrus maxima	GCA_002006925p1_ASM200692v1	1504	345.8	32082701	0	0
Citrus medica	GCA_002013955p2_C_medica_denovo_2	32732	406.1	369527	0	0
Citrus reticulata	GCA_003258625p1_ASM325862v1	67725	344.3	1288159	0	0
Citrus sinensis	Csinensis_154_v1p1pgene_exons	12574	319.2	250548	25379	279876
Citrus sinensis	GCA_000317415p1_Csi_valencia_1p0	4843	327.7	22711823	0	0
Citrus sinensis	GCF_000317415p1_Csi_valencia_1p0	4844	327.8	22711823	28561	35648
Citrus unshiu	GCA_001753815p1_CunshiuBMS10_01	507	1.2	3337	0	0
Citrus unshiu	GCA_002897195p1_CUMW_v1p0	20876	359.7	386404	29039	37970
Citrus x paradisi x Citrus trifoliata	GCA_001929425p1_WD23_11_assembly_v1	238488	265.5	2091	0	0
Coccomyxa sp. LA000219	GCA_000812005p1_ASM81200v1	106	48.5	2254067	0	0
Coccomyxa sp. SUA001	GCA_001244535p1_ASM124453v1	23591	11.8	570	0	0
Coccomyxa subellipsoidea C-169	GCF_000258705p1	29	48.8	1959569	9915	9839
Coelastrella	GCA_001630525p1_ASM163052v1	16225	80.2	9337	0	0
Coelastrella sp. UTEX B 3026	GCA_002588565p1_ASM258856v1	29867	151.5	10705	0	0
Conringia planisiliqua	GCA_900108845p1_Conringia_planisiliquapv1	705	184.2	8882589	0	0
Conyza canadensis	GCA_000775935p1_ASM77593v1	20075	326.2	20748	0	0

Corchorus capsularis	GCA_001974805p1_CCACVL1_1p0	16522	317.2	46451	31069	29356
Corchorus olitorius	GCA_001974825p1_COLO4_1p0	24918	334.9	44998	38582	35704
Cucumis melo	GCA_000313045p1_ASM31304v1	31463	374.8	4278129	0	0
Cucumis melo	GCF_000313045p1_ASM31304v1	31464	374.9	4278129	22741	29798
Cucumis sativus	Csativus_122_v1p0pgene_exons	4219	203.1	993451	21503	177988
Cucumis sativus	GCA_000004075p2_ASM407v2	186	193.8	29076228	23780	23780
Cucumis sativus	GCA_001483825p1_ASM148382v1	6693	278.1	379917	0	0
Cucumis sativus	GCF_000004075p2_ASM407v2	190	195.7	29076228	20405	25668
Cucurbita maxima	GCA_002738345p1_Cmax_1p0	8299	271.4	3717157	0	0
Cucurbita maxima	GCF_002738345p1_Cmax_1p0	8299	271.4	3717157	35289	42777
Cucurbita moschata	GCA_002738365p1_Cmos_1p0	3500	269.9	3995720	0	0
Cucurbita moschata	GCF_002738365p1_Cmos_1p0	3500	269.9	3995720	35355	43715
Cucurbita pepo	GCA_002806865p2_ASM280686v2	25364	260.5	9833969	0	0
Cucurbita pepo	GCF_002806865p1_ASM280686v2	25263	261.4	9833969	35798	43466
Cuscuta australis	GCA_003260385p1_Cau_v1p0	218	262.6	3625894	18157	18157
Cuscuta campestris	GCA_900332095p1_ASM90033209v1	6907	476.8	1384808	0	0
Cyanidioschyzon merolae	GCA_000091205p1_ASM9120v1	20	16.5	859119	6170	4803
Cyanidioschyzon merolae	GCF_000091205p1_ASM9120v1	20	16.5	859119	5373	4803
Cymbomonas tetramitiformis	GCA_001247695p1_ASM124769v1	40243	281.3	10932	0	0
Cynara cardunculus	GCA_001531365p1_CcrdV1	8283	725.2	25947084	26505	26505
Cynara cardunculus	GCF_001531365p1_CcrdV1	8283	725.2	25947084	30288	38406
Dactylis glomerata	GCA_002892645p1_ASM289264v1	1072009	839.9	1656	0	0
Datisca glomerata	GCA_003255025p1_ASM325502v1	13864	688.4	1186304	0	0
Daucus carota	Dcarota_388_v2p0pgene_exons	4826	421.5	36610139	32113	160795
Daucus carota	GCA_001625215p1_ASM162521v1	4826	421.5	36610139	33502	32113
Daucus carota	GCF_001625215p1_ASM162521v1	4826	421.5	36610139	36244	44655
Dendrobium catenatum	GCA_001605985p1_ASM160598v1	72901	1008.5	391462	0	0
Dendrobium catenatum	GCA_001605985p2_ASM160598v2	286089	1104.1	1043725	29149	29149
Dendrobium officinale	GCF_001605985p1_ASM160598v1	72902	1008.7	391462	25123	34527
Dianthus caryophyllus	GCA_000512335p1_DCA_r1p0	45088	567.7	60730	0	0
Dichanthelium oligosanthes	GCA_001633215p2_ASM163321v2	17436	589.2	74581	26468	26468
Dioscorea alata	GCA_002904275p2_ASM290427v2	57706	620.9	19343	0	0
Dioscorea rotundata	C_01_P1_1_P2_18pfinal	21	456.7	25272979	19086	19086

Dioscorea rotundata	GCA_002240015p2_TDr96_F1_Pseudo_Chromosome_v1p0	21	456.7	25272979	0	0
Diospyros lotus	GCA_000774125p1_ASM77412v1	796	1.1	1870	0	0
Dorcoceras hygrometricum	GCA_001598015p1_Boea_hygrometricapv1	401752	1521.4	113694	47778	47778
Drosera capensis	GCA_001925005p1_ASM192500v1	12713	263.8	82649	0	0
Dryas drummondii	GCA_003254865p1_ASM325486v1	13357	225.5	931783	0	0
Dunaliella salina	Dsalina_325_v1p0	2464	329.8	364726	16697	18801
Dunaliella salina	GCA_002284615p1_Dsal_v1p0	5512	343.7	353034	0	0
Durio zibethinus	GCA_002303985p1_Duzib1p0	677	715.2	22724830	0	0
Durio zibethinus	GCF_002303985p1_Duzib1p0	677	715.2	22724830	44795	63007
Echinochloa crus-galli	GCA_900205405p1_ASM90020540v1	4534	1486.6	1802240	0	0
Eichhornia paniculata	GCA_001647135p1_ASM164713v1	40286	571.4	31651	0	0
Elaeis guineensis	GCA_000442705p1_EG5	40060	1535.0	1268079	0	0
Elaeis guineensis	GCA_001672495p1_ASM167249v1	218141	499.0	2579	0	0
Elaeis guineensis	GCF_000442705p1_EG5	40061	1535.2	1268079	30194	39539
Elaeis oleifera	GCA_000441515p1_EO8	26756	1402.7	333109	0	0
Eleusine coracana	GCA_002180455p1_ASM218045v1	525627	1196.0	23733	0	0
Embelia ribes	GCA_001753735p1_Embelia_ribes_ER1_v1	107000	660.5	8704	0	0
Ensete ventricosum	GCA_000818735p2_Ensete_Bedadeti_v2p0	45745	451.3	21097	0	0
Ensete ventricosum	GCA_001884845p1_Onjamo_v1p0	51525	444.8	16208	0	0
Eragrostis tef	GCA_000970635p1_ASM97063v1	13883	607.3	116204	0	0
Erigeron canadensis	GCA_000775935p1_ASM77593v1	20075	326.2	20748	0	0
Erythranthe guttata	GCA_000504015p1_Mimgu1_0	2211	321.6	1123783	27890	29504
Erythranthe guttata	GCF_000504015p1_Mimgu1_0	2212	322.2	1123783	30379	31861
Erythranthe guttata	Mguttatus_256_v2p0	421	304.8	21212587	28140	33573
Eschscholzia californica	GCA_002897215p1_ECA_r1p0	53253	489.1	752971	0	0
Ettlia oleoabundans	GCA_001937085p1_ASM193708v1	7999	59.3	14136	0	0
Eucalyptus camaldulensis	GCA_000260855p1_EUC_r1p0	274001	654.9	4275	0	0
Eucalyptus grandis	Egrandis_297_v2p0pgene_exons	4943	691.3	57472304	36349	239526
Eucalyptus grandis	GCA_000612305p1_Egrandis1_0	4950	691.3	53892272	36779	46920
Eucalyptus grandis	GCF_000612305p1_Egrandis1_0	4951	691.4	53892272	43939	47423
Euclidium syriacum	GCA_900116095p1_Euclidium_syriacumpMPIPZpv1	160	229.2	17487894	0	0
Eudorina sp. 2006-703-Eu-15	GCA_003117195p1_EudorinaFemale_1p0	3180	184.0	564035	0	0

Euphorbia esula	GCA_002919075p1_ASM291907v1	1633094	1124.9	1035	0	0
Eutrema heterophyllum	GCA_002933915p1_ASM293391v1	57686	349.0	561173	0	0
Eutrema salsugineum	Esalsugineum_173_v1p0pgene_exons	639	243.1	13441892	26351	160003
Eutrema salsugineum	GCA_000325905p2_TsV2-8	2155	231.9	25023397	0	0
Eutrema salsugineum	GCA_000478725p1_Eutsalg1_0	638	243.1	13441892	26528	29485
Eutrema salsugineum	GCF_000478725p1_Eutsalg1_0	638	243.1	13441892	33009	33637
Eutrema yunnanense	GCA_002933935p1_ASM293393v1	78020	415.4	371182	0	0
Fagopyrum esculentum	GCA_001661195p1_FES_r1p0	387594	1177.7	25109	0	0
Fagopyrum tataricum	GCA_002319775p1_Ft1p0	7020	505.9	53883329	0	0
Fagus sylvatica	GCA_900244945p1_Beech_Genome	6491	542.3	145397	0	0
Ficus carica	GCA_002002945p1_Fpcarica_assembly01	27995	247.1	166092	0	0
Fragaria iinumae	GCA_000511975p1_FII_r1p1	117822	199.6	3309	0	0
Fragaria nipponica	GCA_000512025p1_FNI_r1p1	215024	206.4	1275	0	0
Fragaria nubicola	GCA_000511995p1_FNU_r1p1	210780	203.7	1291	0	0
Fragaria orientalis	GCA_000517285p1_FOR_r1p1	323163	214.2	722	0	0
Fragaria vesca	Fvesca_226_v1p1pgene_exons	8	206.9	27214541	32831	167270
Fragaria vesca	GCA_000184155p1_FraVesHawaii_1p0	3047	214.2	27879571	0	0
Fragaria vesca	GCF_000184155p1_FraVesHawaii_1p0	3048	214.4	27879571	27843	31387
Fragaria x ananassa	GCA_000511835p1_FAN_r1p1	625966	697.8	2201	0	0
Fraxinus excelsior	GCA_900149125p1_BATG-0p5	89515	867.5	104030	0	0
Galdieria sulphuraria	GCF_000341285p1_ASM34128v1	433	13.7	172322	6723	7174
Gastrodia elata	GCA_002966915p1_ASM296691v1	3768	1061.0	4911943	0	0
Genlisea aurea	GCA_000441915p1_GenAur_1p0	10684	43.4	5786	17685	17685
Geum urbanum	GCA_900236755p1_G_urb_d1	170029	1217.0	24601	0	0
Glycine max	GCA_000004515p4_Glycine_max_v2p1	1190	978.5	48577505	56044	88647
Glycine max	GCF_000004515p4_Glycine_max_v2p0	1191	979.0	48577505	58882	71525
Glycine max	Gmax_275_Wm82pa2pv1pgene_exons	1190	978.5	48577505	56044	525934
Glycine soja	GCA_000722935p2_W05v1p0	33170	863.6	404776	50399	50399
Glycine soja	GCA_002907465p1_glysopPI483463pgnm1	306	985.3	48820272	0	0
Glycyrrhiza uralensis	Gurpdraft-genomep20151208	4853	325.3	133536	34445	38135
Gonium pectorale	GCA_001584585p1_ASM158458v1	2373	148.8	1267136	16290	16290
Gossypioides kirkii	GCA_002818315p1_Gokirpv1	745	528.7	41165770	0	0
Gossypium arboreum	GCA_000612285p2_Gossypium_arboreum_v1p0	75418	1694.4	121339338	0	0

Gossypium arboreum	GCA_000787975p1_arboreum_v1p0	392831	1862.2	22252	39320	33609
Gossypium arboreum	GCF_000612285p1_Gossypium_arboreum_v1p0	75419	1694.6	121339338	40208	47568
Gossypium barbadense	GCA_001856525p1_GbV1p0	29751	2566.7	259869	0	0
Gossypium hirsutum	GCA_000987745p1_ASM98774v1	9146	2188.3	70911690	0	0
Gossypium hirsutum	GCF_000987745p1_ASM98774v1	9148	2189.1	70911690	78218	90927
Gossypium raimondii	GCA_000327365p1_Graimondii2_0	1033	761.4	62175169	38208	78371
Gossypium raimondii	GCA_000331045p1_Gr_v1p0	4699	773.8	2284095	0	0
Gossypium raimondii	GCF_000327365p1_Graimondii2_0	1034	761.6	62175169	44724	59057
Gossypium raimondii	Graimondii_221_v2p1pgene_exons	1033	761.4	62175169	37505	486043
Gracilariopsis chorda	GCA_003194525p1_GraCho1p0	1211	92.2	220274	10938	10806
Gracilariopsis lemaneiformis	GCA_003346895p1_Glem_v01	13775	88.7	34594	0	0
Handroanthus impetiginosus	GCA_002762385p1_Himp0p1	13204	503.3	80946	30271	30271
Helianthus annuus	GCA_002127325p1_HanXRQr1p0	1528	3027.8	178899001	57832	52230
Helianthus annuus	GCF_002127325p1_HanXRQr1p0	1528	3027.8	178899001	81678	73839
Helicosporidium sp. ATCC 50920	GCA_000690575p1_Helico_v1p0	5666	12.4	3036	6033	6033
Herrania umbratica	GCA_002168275p2_ASM216827v2	6132	234.7	8132550	0	0
Herrania umbratica	GCF_002168275p1_ASM216827v2	6074	234.0	8132550	20744	27748
Hevea brasiliensis	GCA_001654055p1_ASM165405v1	7452	1373.4	1281786	0	0
Hevea brasiliensis	GCF_001654055p1_ASM165405v1	7453	1373.5	1281786	42686	58062
Hibiscus syriacus	GCA_001696755p1_ASM169675v1	77488	1748.3	139874	0	0
Hordeum bulbosum	GCA_900070015p1_Hordeum_bulbosum_assembly1	2883554	1294.9	511	0	0
Hordeum pubiflorum	GCA_000582825p1_Hordeum_pubiflorum_assembly1	1818420	1425.3	1662	0	0
Hordeum vulgare	GCA_900075435p2_barley_BACs_2	72295	9788.9	156010	0	0
Hordeum vulgare	Hordeum_vulgarepIBSC_v2p41	10	4834.4	657224000	43051	236301
Hordeum vulgare subsp. vulgare	barley_morex_pseudomolecules	8	4833.8	657224000	0	248180
Hordeum vulgare subsp. vulgare	GCA_000326125p1_ASM32612v1	2077901	1779.5	1986	0	0
Humulus lupulus var. cordifolius	GCA_000830395p1_hl_KR_version_1p0pfasta	132476	2049.2	37081	0	0
Humulus lupulus var. lupulus	GCA_000831365p1_hl_SW_version_1p0pfasta	132476	2049.2	37081	0	0
Ipomoea batatas	GCA_002525835p2_ipoBat4	28461	837.0	41463214	0	0
Ipomoea nil	GCA_001879475p1_Asagao_1p1	3418	735.2	2880368	195	119
Ipomoea nil	GCF_001879475p1_Asagao_1p1	3418	735.2	2880368	47872	51054

Ipomoea trifida	GCA_000978395p1_ITR_r1p0	77400	513.0	42586	0	0
Ipomoea trifida	GCA_000981105p1_ITRk_r1p0	181194	712.2	36283	0	0
Jatropha curcas	GCA_000208675p2_JAT_r4p5	39277	297.7	15950	0	0
Jatropha curcas	GCA_000696525p1_JatCur_1p0	6023	318.4	746835	27172	27172
Jatropha curcas	GCF_000696525p1_JatCur_1p0	6024	318.5	746835	23592	28814
Juglans cathayensis	GCA_003122765p1_ASM312276v1	19972	600.2	193887	0	0
Juglans hindsii	GCA_003123825p1_ASM312382v1	73433	611.1	487794	0	0
Juglans mandshurica	GCA_002916435p1_m4v1	13809	558.1	496923	0	0
Juglans microcarpa	GCA_003123845p1_ASM312384v1	112570	914.0	141324	0	0
Juglans nigra	GCA_003123865p1_ASM312386v1	90472	620.8	252148	0	0
Juglans regia	GCA_001411555p1_wgsp5d	105811	700.6	250485	0	0
Juglans regia	GCF_001411555p1_wgsp5d	105803	699.7	250522	43323	55627
Juglans sigillata	GCA_003123805p1_ASM312380v1	134300	648.1	207533	0	0
Kalanchoe fedtschenkoi	GCA_002312845p1_K_fedtschenkoi_M2_v1	1324	256.4	2451343	0	0
Kalanchoe fedtschenkoi	Kfedtschenkoi_382_v1p1	778	254.2	2451343	30964	45190
Kalanchoe laxiflora	Klaxiflora_309_v1p1pgene_exons	3221	422.0	454876	50461	411261
Kalanchoe laxiflora	Klaxifora_309_v1p0	2120	418.8	457852	50461	69177
Kappaphycus alvarezii	GCA_002205965p2_ASM220596v2	899	336.7	848967	0	0
Klebsormidium flaccidum	GCA_000708835p1_ASM70883v1	1814	104.2	134930	16273	16283
Klebsormidium nitens	GCA_000708835p1_ASM70883v1	1814	104.2	134930	16273	16283
Kokia drynarioides	GCA_002814295p1_KokDry1	15383	517.4	177976	0	0
Lactuca sativa	GCA_000227445p1_Legassy_V2	876110	1133.7	2172	0	0
Lactuca sativa	GCA_002870075p1_Lsat_Salinas_v7	11452	2384.0	1769135	38693	38294
Lactuca sativa	GCF_002870075p1_Lsat_Salinas_v7	11453	2384.2	1769135	46234	45242
Lagenaria siceraria	GCA_000466325p1_Bottle_gourd	305112	176.7	782	0	0
Lagenaria siceraria	GCA_003268545p1_Lsi_v1p0	438	313.4	8701157	0	0
Leavenworthia alabamica	GCA_000411055p1_VEGI_LA_v_1p0	11715	173.4	71084	0	0
Leersia perrieri	GCA_000325765p3_Lperr_V1p4	12	266.7	22540073	0	0
Linum usitatissimum	GCA_000224295p1_LinUsi_v1p1	48397	282.2	21193	0	0
Linum usitatissimum	Lusitatissimum_200_v1p0	1028	293.5	781883	43471	43484
Liriodendron chinense	GCA_003013855p1_ASM301385v1	217583	1561.1	1015738	0	0
Lolium perenne	GCA_001735685p1_ASM173568v1	666180	481.5	967	0	0
Lophocereus schottii	GCA_002740545p1_Lsch_v1p3	158704	797.9	9302	0	0
Lotus japonicus	GCA_000181115p2_Lj3p0	44464	394.5	25054	0	0

Lotus japonicus	Lj3p0	8	447.0	62285374	83083	79471
Lupinus angustifolius	GCA_000338175p1_Lupin	71995	523.3	15485	0	0
Lupinus angustifolius	GCA_001865875p1_LupAngTanjil_v1p0	13573	609.2	21299880	33074	33083
Lupinus angustifolius	GCF_001865875p1_LupAngTanjil_v1p0	13573	609.2	21299880	38688	52821
Macadamia integrifolia	GCA_900087525p1_Macadmia_integrifolia_v1p1	193493	518.5	4745	0	0
Macleaya cordata	GCA_002174775p1_MC_HNAU_1p0	4547	377.8	308204	21911	21911
Malus domestica	GCA_000148765p2_MalDomGD1p0	1250	1874.4	2966274	0	0
Malus domestica	GCF_000148765p1_MalDomGD1p0	1251	1874.8	2966274	58136	60549
Malus domestica	Mdomestica_196_v1p0pgene_exons	122107	881.3	11136	63514	301245
Manihot esculenta	GCA_001659605p1_Manihot_esculenta_v6	2019	582.1	28119335	33044	41393
Manihot esculenta	GCF_001659605p1_Manihot_esculenta_v6	2020	582.3	28119335	31954	43286
Manihot esculenta	Mesculenta_305_v6p1	479	554.8	28438989	33033	41381
Manihot esculenta subsp. flabellifolia	GCA_000737105p1_MW_v2d	54016	390.8	14635	0	0
Marchantia polymorpha	GCA_001641455p1_Mp_v4	4137	205.7	372128	17956	17956
Marchantia polymorpha	GCA_003032435p1_Marchanta_polymorpha_v1	2957	225.8	1366373	19287	24674
Marchantia polymorpha	Mpolymorpha_320_v3p1	763	215.5	1407541	19287	24674
Medicago truncatula	GCA_000219495p2_MedtrA17_4p0	2186	412.8	49172423	51519	57585
Medicago truncatula	GCF_000219495p3_MedtrA17_4p0	2187	412.9	49172423	51628	57661
Medicago truncatula	Mtruncatula_285_Mt4p0v1pgene_exons	1949	411.8	49172423	50894	284973
Melia azedarach	MELAZ155640_EIv1pannotation	550	230.8	3132033	26738	165241
Mentha longifolia	GCA_001642375p1_Mlong1p0	190876	353.3	3645	0	0
Metrosideros polymorpha var. glaberrima	GCA_001662345p1_Mpo_1p0	36376	304.4	5051733	0	0
Micractinium conductrix	GCA_002245815p2_ASM224581v2	300	61.0	1210495	9217	10070
Micromonas commoda	GCF_000090985p2_ASM9098v2	19	21.1	1394110	10127	10140
Micromonas pusilla CCMP1545	GCF_000151265p2	21	22.0	1183541	10248	10242
Micromonas pusilla CCMP1545	MpusillaCCMP1545_228_v3p0	21	21.9	1183541	10660	10660
Micromonas sp. ASP10-01a	GCA_001430725p1_ASM143072v1	1069	19.6	22484	0	0
Micromonas commoda	MspRCC299_229_v3p0	17	21.0	1394110	10103	10103
Mimosa pudica	GCA_003254945p1_ASM325494v1	97892	557.2	119676	0	0
Miscanthus sacchariflorus	GCA_002993905p1_Msac_v3	105321	2074.9	37709	0	0

Momordica charantia	GCA_001995035p1_ASM199503v1	1052	285.6	1100631	0	0
Momordica charantia	GCF_001995035p1_ASM199503v1	1052	285.6	1100631	21684	28666
Monoraphidium neglectum	GCA_000611645p1_mono_v1	6720	69.7	15659	16807	16755
Monoraphidium neglectum	GCF_000611645p1_mono_v1	6720	69.7	15659	16807	16755
Monoraphidium sp. 549	GCA_002814315p1_ASM281431v1	1851	74.7	105989	0	0
Monotropa hypopitys	GCA_002855965p1_monotropa1p0	1259264	2197.5	2546	0	0
Morus notabilis	GCA_000414095p2_ASM41409v2	31301	320.4	405448	29261	26965
Morus notabilis	GCF_000414095p1_ASM41409v2	31301	320.4	405448	29261	26965
Musa acuminata	GCA_000313855p2_ASM31385v2	7259	472.2	28617404	0	0
Musa acuminata	Macuminata_304_v1pgene_exons	12	473.0	34148863	36528	197588
Musa acuminata subsp. malaccensis	GCF_000313855p2_ASM31385v2	7259	472.2	28617404	33417	41734
Musa itinerans	GCA_001649415p1_ASM164941v1	28415	455.3	195772	0	0
Nelumbo nucifera	GCA_000365185p2_Chinese_Lotus_1p1	3619	805.1	3435397	130	84
Nelumbo nucifera	GCA_000805495p1_Nelumbo_nucifera_v1p1	14895	790.3	989329	0	0
Nelumbo nucifera	GCF_000365185p1_Chinese_Lotus_1p1	3603	804.6	3435397	29034	38964
Nicotiana attenuata	GCA_001879085p1_NIATTr2	37194	2365.7	524499	33320	33320
Nicotiana attenuata	GCF_001879085p1_NIATTr2	37194	2365.7	524499	39977	44491
Nicotiana glauca	GCA_002930595p1_NicGla1p0	514289	3222.8	30470	0	0
Nicotiana obtusifolia	GCA_002018475p1_NIOBTpversion3	53128	1222.8	134141	0	0
Nicotiana otophora	GCA_000715115p1_Noto	929607	2689.4	26649	0	0
Nicotiana sylvestris	GCA_000393655p1_Nsyl	253917	2221.8	79726	0	0
Nicotiana sylvestris	GCF_000393655p1_Nsyl	253918	2222.0	79727	41187	48160
Nicotiana tabacum	GCA_000715135p1_Ntab-TN90	351737	3718.8	66158	0	0
Nicotiana tabacum	GCF_000715135p1_Ntab-TN90	168247	3643.5	67743	74273	84255
Nicotiana tomentosiformis	GCA_000390325p2_Ntom_v01	159547	1688.3	82593	0	0
Nicotiana tomentosiformis	GCF_000390325p1_Ntom_v01	159549	1688.5	82598	38190	45607
Nissolia schottii	GCA_003254905p1_ASM325490v1	116213	466.1	179654	0	0
Nothapodytes nimmoniana	GCA_002091855p1_Nnimmo_assembly01	2301	1.4	785	0	0
Ochetophila trinervis	GCA_003254975p1_ASM325497v1	8237	309.1	115526	0	0
Ocimum tenuiflorum	GCA_001278415p1_OciTen1p0	121993	332.6	5674	0	0
Olea europaea	GCA_002742605p1_O_europaea_v1	41225	1141.0	12567911	0	0
Olea europaea	GCA_003313485p1_Duke_Pbarb_2016	5473	1214.8	468024	56349	89982

Olea europaea	GCF_002742605p1_O_europaea_v1	41226	1141.1	12567911	47911	58334
Oropetium thomaeum	GCA_001182835p1_Oropetium	625	243.2	2386382	0	0
Oropetium thomaeum	Othomaeum_386_v1p0pgene_exons	625	243.2	2386382	28446	129424
Oryza barthii	GCA_000182155p3_Opbarthii_v1p3	12	308.3	25711811	0	0
Oryza brachyantha	GCA_000231095p2_Oryza_brachyanthapv1p4b	2491	259.9	21479432	0	0
Oryza brachyantha	GCF_000231095p1_Oryza_brachyanthapv1p4b	2491	259.9	21479432	24828	26803
Oryza glaberrima	GCA_000147395p2_Oryza_glaberrima_V1	25599	303.3	23146	0	0
Oryza glumipatula	GCA_000576495p1_Oryza_glumaepatula_v1p5	12	372.9	31548187	0	0
Oryza longistaminata	GCA_001514335p2_ASM151433v2	9688	362.1	30401905	0	0
Oryza meridionalis	GCA_000338895p2_Oryza_meridionalis_v1p3	12	335.7	30391017	0	0
Oryza nivara	GCA_000576065p1_Oryza_nivara_v1p0	12	338.0	28646061	0	0
Oryza punctata	GCA_000573905p1_Oryza_punctata_v1p2	12	393.8	31244610	0	0
Oryza rufipogon	GCA_000817225p1_OR_W1943	3818	339.2	27785585	0	0
Oryza rufipogon	GCA_001551805p1_ASM155180v1	2582	384.5	219409	0	0
Oryza sativa aus subgroup	GCA_001952365p1_ASM195236v1	12	362.3	29936233	0	0
Oryza sativa	GCA_001433935p1_IRGSP-1p0	55	373.8	29958434	46019	48407
Oryza sativa	Osativa_323_v7p0pgene_exons	14	374.5	29958434	42189	239565
Oryza sativa Indica	GCA_000004655p2_ASM465v1	10490	426.3	31162561	39285	37358
Oryza sativa Indica Group	GCA_001305255p1_ASM130525v1	12	352.1	30903862	0	0
Oryza sativa Indica Group	GCA_001618795p1_ZSv2p0	2300	386.5	31109867	0	0
Oryza sativa Japonica	GCF_001433935p1_IRGSP-1p0	58	374.4	29958434	33189	41070
Oryza sativa Japonica Group	GCA_000321445p1_Osat_hitom_01	12	382.6	31217802	0	0
Ostreococcus lucimarinus CCE9901	GCF_000092065p1_ASM9206v1	21	13.2	708927	7640	7603
Ostreococcus lucimarinus	Olucimarinus_231_v2p0	21	13.2	708927	7796	7796
Ostreococcus sp. 'lucimarinus'	GCF_000092065p1_ASM9206v1	21	13.2	708927	7640	7619
Ostreococcus tauri	GCF_000214015p2_version_050606	22	12.6	739027	8114	7994
Pachycereus pringlei	GCA_002740445p1_Ppri_v1p3	171584	629.7	5411	0	0
Panicum hallii	GCA_002211085p2_PHallii_v3p1	291	535.9	57869027	33805	44192
Panicum hallii	GCF_002211085p1_PHallii_v3p1	291	535.9	57869027	31528	37612
Panicum hallii	Phallii_308_v2p0pgene_exons	8414	554.1	59822759	37232	254657
Panicum miliaceum	GCA_002895445p2_ASM289544v2	466	848.4	48259421	0	0
Panicum virgatum	Pvirgatum_273_v1p1	33649	1271.7	55704564	98007	125439
Parachlorella kessleri	GCA_001598975p1_PK2152_assembly	3651	59.2	33885	0	0

Parasponia andersonii	GCA_002914805p1_PanWU01x14_asm01	2732	475.8	712846	37229	37232
Passiflora edulis	GCA_002156105p1_ASM215610v1	234012	165.7	1311	0	0
Penstemon barbatus	GCA_003313485p1_Duke_Pbarb_2016	18827	696.3	43419	0	0
Penstemon centranthifolius	GCA_000737435p1_ASM73743v1	6761	4.5	752	0	0
Penstemon grinnellii	GCA_000737425p1_ASM73742v1	5523	3.7	780	0	0
Pereskia humboldtii	GCA_002740485p1_Phum_v1p3	126352	414.0	4395	0	0
Persea americana	GCA_002908915p1_Hass1p0	5000	446.8	205885	0	0
Phalaenopsis aphrodite	GCA_003013225p1_ASM301322v1	13732	1025.1	946429	0	0
Phalaenopsis equestris	GCA_001263595p1_ASM126359v1	89583	1064.1	378442	0	0
Phalaenopsis equestris	GCF_001263595p1_ASM126359v1	89584	1064.2	378442	21938	29894
Phalaenopsis hybrid cultivar	GCA_002079205p1_ASM207920v1	149149	2687.7	134284	0	0
Phaseolus coccineus	GCA_003122825p1_UCLA_Phcoc_1p0	192921	371.1	7980	0	0
Phaseolus vulgaris	GCA_000499845p1_PhaVulg1_0	708	521.1	50367376	28134	32720
Phaseolus vulgaris	GCF_000499845p1_PhaVulg1_0	708	521.1	50367376	28134	32720
Phoenix dactylifera	GCF_000413155p1_DPV01	80317	556.5	335289	29558	38989
Physcomitrella patens	GCA_000002425p2_Phypa_V3	357	471.9	17435539	31309	31251
Physcomitrella patens	GCF_000002425p4_Phypa_V3	359	472.1	17435539	23747	48022
Physcomitrella patens	Ppatens_318_v3p3	145	471.5	17435539	32926	87533
Picea glauca	GCA_000411955p5_PG29_v4p1	3033322	24633.1	54661	6522	6445
Picea glauca	GCA_000966675p1_WS77111_V1	3353683	26936.2	49216	0	0
Picea glauca	GCA_001687225p1_SeqCapPg29	222034	258.3	1368	0	0
Picochlorum sp. SENEW3	GCA_000876415p1_ASM87641v1	880	13.4	126215	0	0
Picochlorum sp. 'soloecismus'	GCA_002818215p1_ASM281821v1	38	15.3	724710	0	0
Pinus taeda	GCA_000404065p3_Ptaeda2p0	1760464	22103.6	107038	0	0
Pisum sativum	GCA_003013575p1_ASM301357v1	5449423	4275.9	4610	0	0
Populus euphratica	GCA_000495115p1_PopEup_1p0	9614	495.9	482846	0	0
Populus euphratica	GCF_000495115p1_PopEup_1p0	9615	496.0	482055	36439	49760
Populus trichocarpa	GCA_000002775p3_Pop_tri_v3	1446	434.1	19465461	41335	73012
Populus trichocarpa	GCF_000002775p4_Pop_tri_v3	1447	434.3	19465461	37272	51717
Populus trichocarpa	Ptrichocarpa_210_v3p0	379	423.9	19465461	41335	73013
Porphyra umbilicalis	GCA_002049455p2_P_umbilicalis_v1	2126	87.9	202021	13375	13567
Porphyridium purpureum	GCA_000397085p1_Porphyridium_purpureum	3014	19.5	20534	0	0
Primula veris	GCA_000788445p1_ASM78844v1	8756	309.7	165836	0	0
Primula vulgaris	GCA_001077355p1_ASM107735v1	229	1.5	23713	0	0

Primula vulgaris	Pvulgaris_442_v2p1pgene_exons	478	537.2	49670989	27433	218847
Prototheca cutis	GCA_002897115p1_JCM_15793_assembly_v001	29	20.0	1409608	0	0
Prototheca stagnorum	GCA_002794665p1_JCM_9641_assembly_v001	27	16.9	1107247	0	0
Prototheca wickerhamii	GCA_003255715p1_ASM325571v1	3774	27.7	31154	0	0
Prunus avium	GCA_002207925p1_PAV_r1p0	10148	272.4	219566	0	0
Prunus avium	GCF_002207925p1_PAV_r1p0	10148	272.4	219566	30405	35009
Prunus mume	GCA_000346735p1_Ppmume_V1p0	8163	233.9	24358521	0	0
Prunus mume	GCF_000346735p1_Ppmume_V1p0	8164	234.0	24358521	26522	29705
Prunus persica	GCA_000218175p1_PrunusPersicaDD_1p0	30834	214.2	49168	0	0
Prunus persica	GCA_000346465p2_Prunus_persica_NCBIv2	191	227.4	27368013	26873	47089
Prunus persica	GCF_000346465p2_Prunus_persica_NCBIv2	192	227.6	27368013	26412	32595
Prunus persica	Ppersica_298_v2p1	43	226.4	27368013	26873	47089
Prunus yedoensis	GCA_900382725p1_Pynpv1	4016	319.2	145140	0	0
Pseudotsuga menziesii	GCA_001517045p1_DougFir1p0	1236665	14673.2	381586	0	0
Psidium guajava	GCA_002914565p1_Guava1p0	4728	386.9	129242	0	0
Pterocarya stenoptera	GCA_003123785p1_ASM312378v1	124315	955.6	155468	0	0
Punica granatum	GCA_002201585p1_ASM220158v1	17405	296.4	2303557	29226	29226
Purshia tridentata	GCA_003254885p1_ASM325488v1	9353	176.0	33921	0	0
Pyrus x bretschneideri	GCA_000315295p1_Pbr_v1p0	2182	508.6	535028	0	0
Pyrus x bretschneideri	GCF_000315295p1_Pbr_v1p0	2182	508.6	535028	42180	46174
Quercus lobata	GCA_001633185p1_ValleyOak0p5	40156	759.2	95130	0	0
Quercus robur	GCA_900291515p1_Q_robur_v1	550	814.3	55068941	0	0
Quercus suber	GCA_002906115p1_CorkOak1p0	23344	953.3	465160	79750	83282
Quercus suber	GCF_002906115p1_CorkOak1p0	23344	953.3	465160	58326	59614
Quillaja saponaria	GCA_003338715p1_DraftpQuillajapv1p0	48349	248.9	6076	0	0
Quillaja saponaria	QUISA32244_EIv1pannotation	769	354.9	5518683	36027	221643
Raphanus raphanistrum	GCA_000769845p1_ASM76984v1	64732	253.8	10186	0	0
Raphanus sativus	GCA_000801105p2_Rs1p0	10674	426.2	38354807	0	0
Raphanus sativus	GCF_000801105p1_Rs1p0	10676	426.6	38354807	58745	61216
Raphidocelis subcapitata	GCA_003203535p1_Rsub_1p0	300	51.2	341804	13429	13383
Rhazya stricta	GCA_001752375p1_RHA1p0	979	274.4	5553863	0	0
Rhizophora apiculata	GCA_900174605p1_Rap_scaffold_v2	142	232.1	5420131	0	0

Ricinus communis	GCA_000151685p2_JCVI_RCG_1p1	25763	350.6	496528	32025	31307
Ricinus communis	GCF_000151685p1_JCVI_RCG_1p1	25763	350.6	496528	22334	27998
Ricinus communis	Rcommunis_119_v0p1pgene_exons	25828	350.6	496528	31221	129291
Rosa chinensis	GCA_002994745p1_RchiOBHm-V2	47	514.3	69643165	50539	45466
Rosa chinensis	GCF_002994745p1_RchiOBHm-V2	45	513.9	69643165	40349	44948
Rosa multiflora	GCA_002564525p1_RMU_r2p0	83189	739.6	90830	0	0
Rosa x damascena	GCA_001662545p1_ASM166254v1	307872	711.7	27573	0	0
Ruellia speciosa	GCA_001909325p1_Rspec1p0	794288	740.0	1201	0	0
Saccharum hybrid cultivar	GCA_900465005p1_MTP	5708	530.7	116672	0	0
Saccharum spontaneum	GCA_900500655p1_Sugarcane	75981	3924.2	89080	0	0
Salix purpurea	Spurpurea_289_v1p0	2780	450.1	17358976	37865	61520
Salvia miltiorrhiza	Salvia_miltiorrhiza_manual_add	9355	420.0	63197	30478	30478
Santalum album	GCA_002911635p1_ASM291163v1	180	196.1	4363285	0	0
Scenedesmus quadricauda	GCA_002317545p1_ASM231754v1	13425	65.4	8094	0	0
Schrenkiella parvula	GCA_000218505p1_Eutrema_parvulum_v01	1457	137.1	16150104	0	0
Secale cereale	GCA_900079665p1_Rye_Lo7_WGS_contigs	1581707	1684.9	1708	0	0
Selaginella kraussiana	GCA_001021135p1_ASM102113v1	105914	114.5	2415	0	0
Selaginella moellendorffii	GCA_000143415p2_v1p0	758	212.5	1749879	34782	34807
Selaginella moellendorffii	GCF_000143415p4_v1p0	757	212.3	1749879	37888	45247
Selaginella moellendorffii	Smoellendorffii_91_v1p0pgene_exons	768	212.8	1749879	22285	122857
Selaginella tamariscina	GCA_003024785p1_ASM302478v1	1391	300.7	407666	0	0
Sesamum indicum	GCA_000512975p1_S_indicum_v1p0	16235	274.9	17356267	0	0
Sesamum indicum	GCA_001692995p1_S_indicum_Yuzhi11_v1	5868	210.8	324903	0	0
Sesamum indicum	GCF_000512975p1_S_indicum_v1p0	16236	275.1	17356267	26123	33093
Setaria italica	GCA_000263155p2_Setaria_italica_v2p0	336	405.7	47252588	34584	43001
Setaria italica	GCA_001652605p1_ASM165260v1	2689	477.5	53212001	0	0
Setaria italica	GCF_000263155p2_Setaria_italica_v2p0	337	405.9	47252588	31102	32964
Setaria italica	Sitalica_312_v2p2pgene_exons	336	405.7	47253416	34584	218186
Setaria viridis	Sviridis_311_v1p1	130	392.8	46083338	35214	48594
Silene latifolia	GCA_003260165p1_S_latifolia_v1p0	319506	1185.1	10814	0	0
Silene latifolia subsp. alba	GCA_001412135p1_ASM141213v1	307720	665.3	3519	0	0
Silybum marianum	GCA_001541825p1_ASM154182v1	258575	1477.6	6967	0	0
Sisymbrium irio	GCA_000411075p1_VEGI_SI_v_1p0	21357	245.6	144321	0	0
Solanum americanum	GCA_900188915p1	837	9.0	10942	0	0

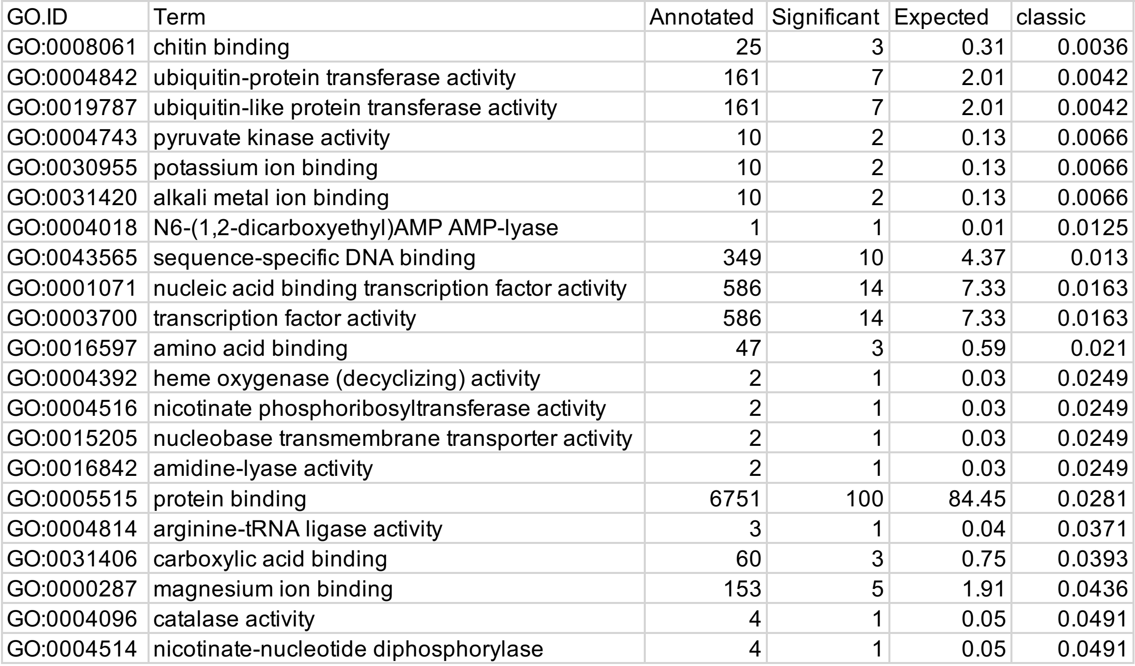
Solanum arcanum	GCA_000612985p1_Soarc10	46594	665.2	31288	0	0
Solanum commersonii	GCA_001239805p1_ASM123980v1	63664	729.6	38514	0	0
Solanum habrochaites	GCA_000577655p1_Sohab10	42990	724.3	37085	0	0
Solanum lycopersicum	GCA_900008105p1_V100	13	760.1	64845585	0	0
Solanum lycopersicum	GCF_000188115p3_SL2p50	3145	823.8	66470942	31075	36149
Solanum lycopersicum	Slycopersicum_390_ITAG2p4pgene_exons	13	823.9	66470942	34725	157233
Solanum melongena	GCA_000787875p1_SME_r2p5p1	33873	833.1	64530	0	0
Solanum pennellii	GCA_001406875p2_SPENNV200	13	926.6	77991103	0	0
Solanum pennellii	GCF_001406875p1_SPENNV200	12	926.4	77991103	32519	35068
Solanum pimpinellifolium	GCA_000230315p1_Sol_pimpi_v1p0	309180	688.2	5714	0	0
Solanum tuberosum	GCA_000226075p1_SolTub_3p0	14853	705.8	1344915	0	0
Solanum tuberosum	GCF_000226075p1_SolTub_3p0	14854	705.9	1344915	33606	37966
Solanum tuberosum	Stuberosum_206_v3p4	12	705.9	61095886	35119	51472
Solanum tuberosum	Stuberosum_448_v4p03pgene_exons	13	773.0	59756223	39028	202449
Solanum verrucosum	GCA_900185145p1_discovar-mp-dt-bn	224100	730.1	4584101	0	0
Sorghum bicolor	GCA_000003195p3_Sorghum_bicolor_NCBIv3	867	708.7	68658214	34118	47110
Sorghum bicolor	GCF_000003195p3_Sorghum_bicolor_NCBIv3	869	709.3	68658214	32945	39248
Sorghum bicolor	Sbicolor_313_v3p1	94	711.0	68658214	34211	47205
Sphagnum fallax	Sfallax_310_v0p5pgene_exons	1228	396.4	1834521	26939	187681
Spinacia oleracea	GCA_000510995p2_Spinach-1p0p3	103502	493.8	19014	21540	23522
Spinacia oleracea	GCA_002007265p1_ASM200726v1	78262	869.8	319471	0	0
Spinacia oleracea	GCF_002007265p1_ASM200726v1	78263	869.9	319471	31764	32794
Spirodela polyrhiza	GCA_001981405p1_ASM198140v1	20	136.7	7641483	0	0
Spirodela polyrhiza	Spolyrhiza_290_v2	33	145.2	4924802	19623	19623
Stenocereus thurberi	GCA_002740465p1_Sthu_v1p3	159477	853.3	10456	0	0
Tarenaya hassleriana	GCA_000463585p1_ASM46358v1	12249	249.9	1600628	0	0
Tarenaya hassleriana	GCF_000463585p1_ASM46358v1	12249	249.9	1600628	30032	40658
Tetrabaena socialis	GCA_002891735p1_TetSoc1	5856	135.8	145927	14296	14296
Tetradesmus obliquus	GCA_900108755p1_sob1	1368	107.7	186615	0	0
Thellungiella parvula	TpV84_ORFs_edit	2136	123.6	6763654	0	141785
Theobroma cacao	GCA_000208745p2_Criollo_cocoa	430	324.7	36364294	0	0
Theobroma cacao	GCA_000403535p1	711	346.0	34397752	29234	44186
Theobroma cacao	GCF_000208745p1_Criollo_cocoa	431	324.9	36364294	24957	30854
Theobroma cacao	Tcacao_233_v1p1pgene_exons	713	346.2	34397752	29452	264870

Thlaspi arvense	GCA_000956625p1_T_arvense_v1	6768	343.0	140815	0	0
Trebouxia gelatinosa	GCA_000818905p1_ASM81890v1	848	61.7	3512598	0	0
Trebouxia sp. TZW2008	GCA_002118135p1_TrTZW2008_1p0	677	69.3	223445	0	0
Trema orientalis	GCA_002914845p1_TorRG33x02_asm01	2756	388.0	656203	35849	35852
Trifolium pratense	GCA_900079335p1_Trpr	39051	346.0	22682783	0	0
Trifolium pratense	Tpratense_385_v2pgene_exons	39051	346.0	22682783	39948	179274
Trifolium subterraneum	GCA_001742945p1_TSUd_r1p1	27424	471.8	287605	42704	42059
Triticum aestivum	GCA_900067645p1	735943	13427.4	88778	0	0
Triticum aestivum	GCA_900241085p1_wheat_TGACv2	519179	13916.9	285110	0	0
Triticum aestivum	IWGSC_v1p1_HC_20170706	22	14547.3	709773743	107891	713422
Triticum aestivum	Taestivum_296_v2p2	86710	634.4	11402	99386	293053
Triticum dicoccoides	GCA_900184675p1_WEW_v1	149145	10495.0	726427787	0	0
Triticum dicoccoides	151210_zavitan_WEW_v2	149145	10509.9	726427787	1686510	1686510
Triticum urartu	GCA_000347455p1_ASM34745v1	499221	3747.0	85725	32265	24169
Triticum urartu	GCA_003073215p1_Tu2p0	10284	4851.9	661480603	0	0
Urochloa ruziziensis	GCA_003016355p1_Bruz	102577	732.5	27770	0	0
Utricularia gibba	GCA_002189035p1_U_gibba_v2	518	100.7	3446356	0	0
Vaccinium macrocarpon	GCA_000775335p1_ASM77533v1	200203	414.6	4291	0	0
Vaccinium macrocarpon	GCA_000775335p2_ASM77533v2	200203	414.6	4291	0	0
Vicia faba	GCA_001375635p1_VfEP_Reference-Unigene	74659	80.4	1723	0	0
Vigna angularis	GCA_001190045p1_Vigan1p1	37373	466.7	34671004	34180	34172
Vigna angularis	GCA_001723775p1_ASM172377v1	3387	444.4	8174047	0	0
Vigna angularis	GCF_001190045p1_Vigan1p1	37375	467.3	31747250	29523	37769
Vigna radiata	GCA_000741045p2_Vradiata_ver6	2497	463.1	25360630	0	0
Vigna radiata	GCF_000741045p1_Vradiata_ver6	2499	463.6	25360630	29146	35143
Vigna radiata var. radiata	GCA_001584445p1_ASM158444v1	2418	454.9	683756	0	0
Vigna unguiculata subsp. unguiculata	GCA_001687525p1_Cowpea_0p03	224035	695.0	7412	0	0
Viola pubescens	GCA_002752925p1_violet_k79	157716	318.4	3500	0	0
Vitis aestivalis	GCA_001562795p1_VitisNorton_MSU1p0	756125	432.8	772	0	0
Vitis cinerea x Vitis riparia	GCA_001282645p1_BoeWGS1p0	210444	539.6	4127	0	0
Vitis vinifera	GCA_000003745p2_12X	1911	485.3	22385789	26346	26346
Vitis vinifera	GCF_000003745p3_12X	1907	486.2	22385789	28982	38120
Vitis vinifera	Vvinifera_145_Genoscopep12Xpgene_exons	33	486.2	23006712	26346	156765

Volvox carteri	GCA_000143455p1_v1p0	1251	137.7	1491501	14437	14439
Volvox carteri	GCF_000143455p1_v1p0	1251	137.7	1491501	14437	14436
Volvox carteri	Vcarteri_317_v2p1	200	130.2	2599759	14247	16075
Xerophyta viscosa	GCA_002076135p1_ASM207613v1	896	295.5	1670317	0	0
Yamagishiella unicocca	GCA_003116995p1_YamagishiellaPlus_1p0	1461	134.2	666310	0	0
Zea mays	GCA_000005005p6_B73_RefGen_v4	265	2134.4	223902240	39320	131270
Zea mays	GCF_000005005p1_B73_RefGen_v3	523	2067.6	217928451	52500	58290
Zea mays	GCF_000005005p2_B73_RefGen_v4	267	2135.1	223902240	49339	58411
Zea mays	Zmays_284_Ensembl-18_2010-01-MaizeSequencepgene_exons	523	2067.9	217959525	63480	342056
Zea mays	ZmaysPH207_443_v1p1pgene_exons	43291	2156.2	215148664	40557	197789
Zea mays subsp. mays	GCA_001644905p2_Zm-W22-REFERENCE-NRGENE-2p0	191	2133.9	222590201	0	0
Zizania latifolia	GCA_000418225p1_Zizania_latifolia_v01	4522	604.0	604864	0	0
Ziziphus jujuba	GCA_000826755p1_ZizJuj_1p1	4789	437.8	25259912	0	0
Ziziphus jujuba	GCA_001835785p1_ASM183578v1	36119	351.1	754884	0	0
Ziziphus jujuba	GCF_000826755p1_ZizJuj_1p1	4789	437.8	25259912	33324	37526
Zostera marina	GCA_001185155p1_Zosma_marinapvp2p1	2228	203.9	485578	20859	20682
Zostera marina	Zmarina_324_v2p2pgene_exons	2228	203.9	485578	20450	106110
Zoysia japonica	GCA_001602275p1_ASM160227v1	11786	334.4	2370062	0	0
Zoysia matrella	GCA_001602295p1_ASM160229v1	13609	563.4	108897	0	0
Zoysia pacifica	GCA_001602315p1_ASM160231v1	11428	397.0	111449	0	0

Table A2 **GO term enrichment for *Avena strigosa* genes with MITE-like sequences in putative promotor regions**.

Column labelled ‘classic’ denotes *p*-value from hypergeometric test with Bonferonni correction.



GO.ID	Term	Annotated	Significant	Expected	classic
GO:0008061	chitin binding	25	3	0.31	0.0036
GO:0004842	ubiquitin-protein transferase activity	161	7	2.01	0.0042
GO:0019787	ubiquitin-like protein transferase activity	161	7	2.01	0.0042
GO:0004743	pyruvate kinase activity	10	2	0.13	0.0066
GO:0030955	potassium ion binding	10	2	0.13	0.0066
GO:0031420	alkali metal ion binding	10	2	0.13	0.0066
GO:0004018	N6-(1,2-dicarboxyethyl)AMP AMP-lyase	1	1	0.01	0.0125
GO:0043565	sequence-specific DNA binding	349	10	4.37	0.013
GO:0001071	nucleic acid binding transcription factor activity	586	14	7.33	0.0163
GO:0003700	transcription factor activity	586	14	7.33	0.0163
GO:0016597	amino acid binding	47	3	0.59	0.021
GO:0004392	heme oxygenase (decyclizing) activity	2	1	0.03	0.0249
GO:0004516	nicotinate phosphoribosyltransferase activity	2	1	0.03	0.0249
GO:0015205	nucleobase transmembrane transporter activity	2	1	0.03	0.0249
GO:0016842	amidine-lyase activity	2	1	0.03	0.0249
GO:0005515	protein binding	6751	100	84.45	0.0281
GO:0004814	arginine-tRNA ligase activity	3	1	0.04	0.0371
GO:0031406	carboxylic acid binding	60	3	0.75	0.0393
GO:0000287	magnesium ion binding	153	5	1.91	0.0436
GO:0004096	catalase activity	4	1	0.05	0.0491
GO:0004514	nicotinate-nucleotide diphosphorylase	4	1	0.05	0.0491