SYSTEM-BASED TRANSCRIPTOMIC AND METABOLOMIC NETWORK ANALYSES IN *PAPAVER SOMNIFERUM*

by

JOHN WILLIAM KERRY

(Under the Direction of Jim Leebens-Mack)

ABSTRACT

Papaver somniferum is an annual plant which produces economically important medicinal alkaloids, oils, and seeds. The morphine biosynthesis pathway in *P. somniferum* has been characterized but a system-wide analysis of transcriptional modules that regulate the concentration of alkaloids within the plant has not yet been performed.

In this study we use weighted gene co-expression network analysis (WGCNA) of a de novo assembly of 5 tissues of *P. somniferum*. We use differentially expressed transcripts to build a weighted gene co-expression network using the WGCNA bioconductor package. Finally we analyze those modules which correlate well with LC-MS/MS generated metabolite concentration values.

INDEX WORDS: *Papaver* Metabolite Co-expression RNASeq

SYSTEM-BASED TRANSCRIPTOMIC AND METABOLOMIC NETWORK ANALYSES IN *PAPAVER SOMNIFERUM*

by

JOHN WILLIAM KERRY

BS, University of Georgia, 2011

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

SYSTEM-BASED TRANSCRIPTOMIC AND METABOLOMIC NETWORK ANALYSES IN *PAPAVER SOMNIFERUM*

by

JOHN WILLIAM KERRY

Major Professor: Jim Leebens-mack

Committee: Michelle Momany
Thiab Taha

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

DEDICATION

This thesis is dedicated to Bill Kerry as his legacy comes to fruition. This work would not have been possible without his support and love. This thesis is also dedicated to my family, Barbara, Bob, and Victoria whom have supported my studies and provided all the love and assistance I could want and more. I would also like to dedicate this work to Donald Wisdom because sometimes we have the chance to change the trajectory of someone's life in an instant and rarely do we see the fruits of those decisions. Lastly and most significantly I would like to dedicate this work to my wife and best friend Roxanne and our son William for providing everything I could want in life and an incredible amount of support.

ACKNOWLEDGEMENTS

I would like to acknowledge the work of Toni Kutchan and Megan Rolf in initial sample preparation, LC-MS/MS analysis, and RNA extraction. I would also like to acknowledge the Leebens-Mack lab for their support, specifically Charlotte Carrigan and Saravanaraj Ayyampalayam for their integral involvement in all of our projects and Alex Harkess for the assembly protocol used in this study. Finally I'd like to acknowledge my committee: Jim Leebens-Mack, Michelle Momany, and Thiab Taha for their continual support during my studies of computational biology and high throughput analysis.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Papaver somniferum is a medically important annual herbaceous plant found in the northern hemisphere that is endemic to southern Europe and Asia. The medical and economic importance of *P. somniferum* derive from its unique profile of benzylisoquineoline alkaloids (BIAs) that are produced uniquely in *P. somniferum* or produced commonly among members of the family Papaveraceae. *P. somniferum* and its sister species *Papaver setigerum* remain the only sources for the analgesic secondary metabolites codeine and morphine. Sanguinarine is another BIA found in *P. somniferum* and that is used as an antimicrobial additive and has been found to have antiproliferative effects on skin cancer cells (Ahmad, Gupta, Husain, Heiskanen, & Mukhtar, 2000).

These secondary metabolite products and more provide a strong economic incentive to elucidate the details of the BIA biosynthetic pathways. Here we take a combined transcriptomic and targeted metabolomic approach to identify transcripts which are co-expressed and may have a regulative effect on BIA concentrations in vivo. For this analysis we focus on those genes that appear to have expression levels similar to the concentrations of the BIAs thebaine, codeine, and morphine.

Bensylisoquineoline Alkaloid Biosynthesis Pathways

The BIAs share the same common initial biosynthetic pathway which consumes two units of L-Tyrosine. One unit is metabolized into dopamine via L-dopa or tyramine, which is catalyzed by dopa decarboxylase or tyrosine decarboxylase respectively (TYDC) (Facchini & De Luca,

1994). The first L-tyrosine unit is converted to 4-hydroxyphenylpyruvate and then converted to 4-hydroxyphenylacetaldehyde (4-HPAA) by tyrosine aminotransferase (TyrAT) (Lee & Facchini, 2011). 4-HPAA and dopamine are processed by norcoclaurine synthase (NCS) to form (S)-norcoclaurine (Lee & Facchini, 2010). (s)-norcoclaurine is methylated sequentially by norcoclaurine 6-O-methyltransferase (6OMT) and coclaurine N-methyltransferase (CNMT) to form (S)-coclaurine and subsequently (S)-N-methylcoclaurine (Choi, Morishige, Shitan, Yazaki, & Sato, 2002; Morishige, Tsujita, Yamada, & Sato, 2000). (S)-N-methylcoclaurine is then converted to (S)-3'-hydroxy-N-methylcoclaurine through a 3'-hydroxylation reaction driven by N-methylcoclaurine 3'-hydroxylase (NMCH) (Pauli & Kutchan, 1998). Finally (S)-Reticuline is produced by the conversion of (S)-3'-hydroxy-N-methylcoclaurine by the enzyme 3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase (4OMT)(Morishige et al., 2000). A detailed view of this biosynthesis pathway can be found in figure 1.

(S)-Reticuline is particularly important in the biosynthesis of BIAs in Papaver somniferum because it is the point at which the majority of the economically important biosynthesis pathways diverge. The secondary metabolite products downstream of (S)-reticuline that are the focus for this study are separated into 2 major groups: The morphinans and the benzophenanthridines. The morphinans include the secondary metabolites thebaine, oripavine, codeine, and morphine. The benzophenanthridines include the secondary metabolites protopine and sanguinarine. An overview of these biosynthesis pathways can be found in Figure 2.

Morphinan biosynthesis pathways ramify after (S)-Reticuline synthesis with one path forming 1,2-dehydroreticulinium through the action of 1,2-dehydroreticulinium synthase (DRS)(Hirata, Poeaknapo, Schmidt, & Zenk, 2004). (R)-reticuline is then produced by 1,2-dehydroreticulinium reductase (DRR)(De-Eknamkul & Zenk, 1992). The enzyme-coding gene

sequences for DRS and DRR have not yet been identified. (R)-reticuline is synthesized to salutaridine and then reduced to salutaridinol by salutaridine synthase (SalSyn)(Gesell et al., 2009) and salutaridine reductase (SalR)(Ziegler et al., 2006) respectively. Salutaridinol is converted by salutaridinol 7-O-acetyltransferase (SalAT)(Grothe, Lenz, & Kutchan, 2001) to 7-O-acetylsalutaridinol. Thebaine is formed from 7-O-acetylsalutaridinol spontaneously(Lenz & Zenk, 1994) or by interaction with the enzyme thebaine synthase (THS) (Fisinger, Grobe, & Zenk, 2007). Thebaine is converted to neopinone by thebaine 6-O-demethylase (T6ODM) (Hagel & Facchini, 2010)which spontaneously forms codeinone or it is converted to oripavine by codeine demethylase (CODM). The reaction to form morphinone from oripavine is catalyzed by T6ODM and can be consumed to form morphine with the assistance of codeinone reductase (COR). Codeinone is converted to codeine through interaction with COR and then can be demethylated to form morphine via CODM (Hagel & Facchini, 2010).

The benzophenanthridine biosynthesis branch begins with conversion of (S)-reticuline to (S)-scoulerine by reaction with the berberine bridge enzyme (Dittrich & Kutchan, 1991). The cytacrome p450 enzymes (S)-chelanthifoline synthase (CheSyn) and (S)-stylopine synthase (StySyn) convert (S)-scoulerine to the intermediate metabolite (S)-cheilanthifoline and then to S-stylopine in two steps (Bauer & H. Zenk, 1991). The reaction between S-stylopine and tetrahydroprotoberberine cis-N-methyltransferase (TNMT) produces (S)-cis-N-methylstylopine (Liscombe & Facchini, 2007). The interaction of (S)-cis-N-methylstylopine and the cytochrome p450 enzymes protopine 6-hydroxylase (P6H) and methylstylopine 14-hydroxylase (MSH) produce 6-hydroxyprotopine with protopine as an intermediate metabolite(Takemura, Ikezawa, Iwasa, & Sato, 2012). 6-hydroxyprotopine spontaneously rearranges to form

dihydrosanguinarine and is oxidized to sanguinarine by dihydropbenzophenanthridine oxidase (DBOX) (Hagel et al., 2012).

The genes coding for enzymes in the morphinan branch of the BIA biosynthesis pathways have largely been identified, isolated, and cloned. Information such as gene copy number and a profile of splice variants for these enzyme coding genes is not yet known. Information on transcription factors and other genes that have a regulatory effect on the morphinan biosynthesis pathway is also of interest. A complete system level analysis is required to shed light on these questions. Next generation sequencing applied to RNA-Seq analysis is a cost effective and data rich method to answer all of these questions at a whole system resolution in non-model species.

Pearson Correlation Coefficients and Weighted Gene Co-expression Analysis

The 'guilt by association' principle is defined as the notion that genes involved in a shared biological process will tend to be co-regulated. This implies that the converse is also likely true; Genes that co-express are likely to be involved in the same biological process. This principle extends beyond similar transcriptional regulation to include metabolite regulation and phenotypic traits (Saito, Hirai, & Yonekura-Sakakibara, 2008). There are many ways to assess the notion of Co-Expression in transcriptomics data. Some of these include hierarchical clustering analysis (HCA), principal component analysis (PCA), and batch learning self-organizing map (BL-SOM). The Pearson correlation coefficient (PCC) is also commonly used as a similarity measurement to construct co-expression networks.

Each of these methods has subtle problems that should be addressed when analyzing large transcriptome data sets at the system level. HCA and PCA lack the power to sufficiently separate transcripts at a biological process level. BL-SOM has the power to do this but there is no concrete definition of what constitutes a representative transcriptional unit (DiLeo, Strahan, den Bakker, & Hoekenga, 2011). Direct estimation and implementation of thresholds for PCC values can tend to be overly conservative and potential connections can be missed. A weighted adjacency measure based on a scale-free topology better predicts biological interactions between genes (Zhang & Horvath, 2005). Weighted gene co-expression network analysis, or weighted correlation network analysis, (WGCNA) has been show to be a robust tool for constructing and analyzing correlation networks in highly multivariate and multidimensional data (Langfelder & Horvath, 2008).

WGCNA offers many benefits over competing network construction approaches. WGCNA provides a soft notion of module membership so genes that influence more than one module can be easily identified. A robust and biologically sound measure of adjacency called Topological Overlap is used instead of placing a threshold on pearson correlation coefficients. Topological overlap is based on overlap of network neighbors and is preferable to correlation threshold-based clusters (A. Li & Horvath, 2007). Finally WGCNA utilizes the notion of eigengenes. Eigengenes are defined as the first principal component of the transcription patterns of a given module. The use of eigengenes to define module transcriptional profile makes the process of relating whole classifications of transcripts to metabolite concentrations easy and understandable (figure 1.3 shows the relationship between a module's transcriptional profile and the module eigengene)(Langfelder & Horvath, 2008).

Co-Expression Analyses in Plants and the Use of De Novo Transcriptomics

The utilization of transcriptomics and metabolomics together is a natural choice for identifying genes that are enzymatically involved in BIA synthesis or effect BIA concentrations by other means. The use of the integration of co-expression and metabolite data is an increasingly common analysis approach in model and non-model plant species (Higashi & Saito, 2013; Movahedi, Van Bel, Heyndrickx, & Vandepoele, 2012; Schilmiller, Pichersky, & Last, 2012). Most of the enzyme coding genes in the morphinans and benzophenanthridines have been identified but a few still evade identification such as the previously mentioned 1,2-dehydroreticulinium synthase and reductase enzymes. Recently the PsWRKY transcription factor was implicated in BIA synthesis regulation in response to wounding (Mishra et al., 2013). The identification of novel enzyme coding genes effecting BIA metabolite concentration regulation remains an enticing and economically relevant target for analysis given the possibilities of next generation sequencing.

Co-expression analyses coupled with targeted metabolite analysis is frequently observed in model organisms due to the existence of established microarray tools. These analyses have been particularly effective in Arabidopsis thaliana for identifying metabolomics responses to stress conditions (Caldana et al., 2011) and to identify novel associations between genes and a complete flavonol profile (Yonekura-Sakakibara et al., 2008). Similar work has been done in tomato concerning metabolite regulation during fruit development (Alba et al., 2005).

The application of next generation sequencing to the construction and analysis of co-expression network analysis in non-model organisms is a natural extension of the work done in arabidopsis and tomato. However this approach has some significant statistical and informatics challenges not found with microarray based co-expression analysis. Much care should be taken

to minimize the influence of sequencing errors in the final assembly and the choice of normalization approach needs to be carefully considered (Gongora-Castillo & Buell, 2013). The application of next generation sequencing based transcriptomics to co-expression network analysis in *Papaver somniferum*, however, is not without precedent. Desgagné-Penix et al. have successfully applied 454 pyrosequencing based co-expression analysis of *Papaver somniferum* cultivars with varied metabolite profiles to identify genes associated with BIAs not considered in this study (Desgagné-Penix, Farrow, Cram, Nowak, & Facchini, 2012).

This work illustrates the use of de novo transcriptome assembly of five tissues of *Papaver somniferum* over three biological replicate plants. We test two quality control measures to reduce the cumulative effect of sequencing errors on our assembled contiguous sequences. We use estimated abundance values to construct a system-wide transcriptome co-expression profile and relate the resulting transcriptional modules to measured metabolite concentrations. Finally we consider the GOSlim functional distributions for each module, the distribution of transcripts similar to known BIA biosynthesis enzymes, and the distribution of potential transcription factors across the network.

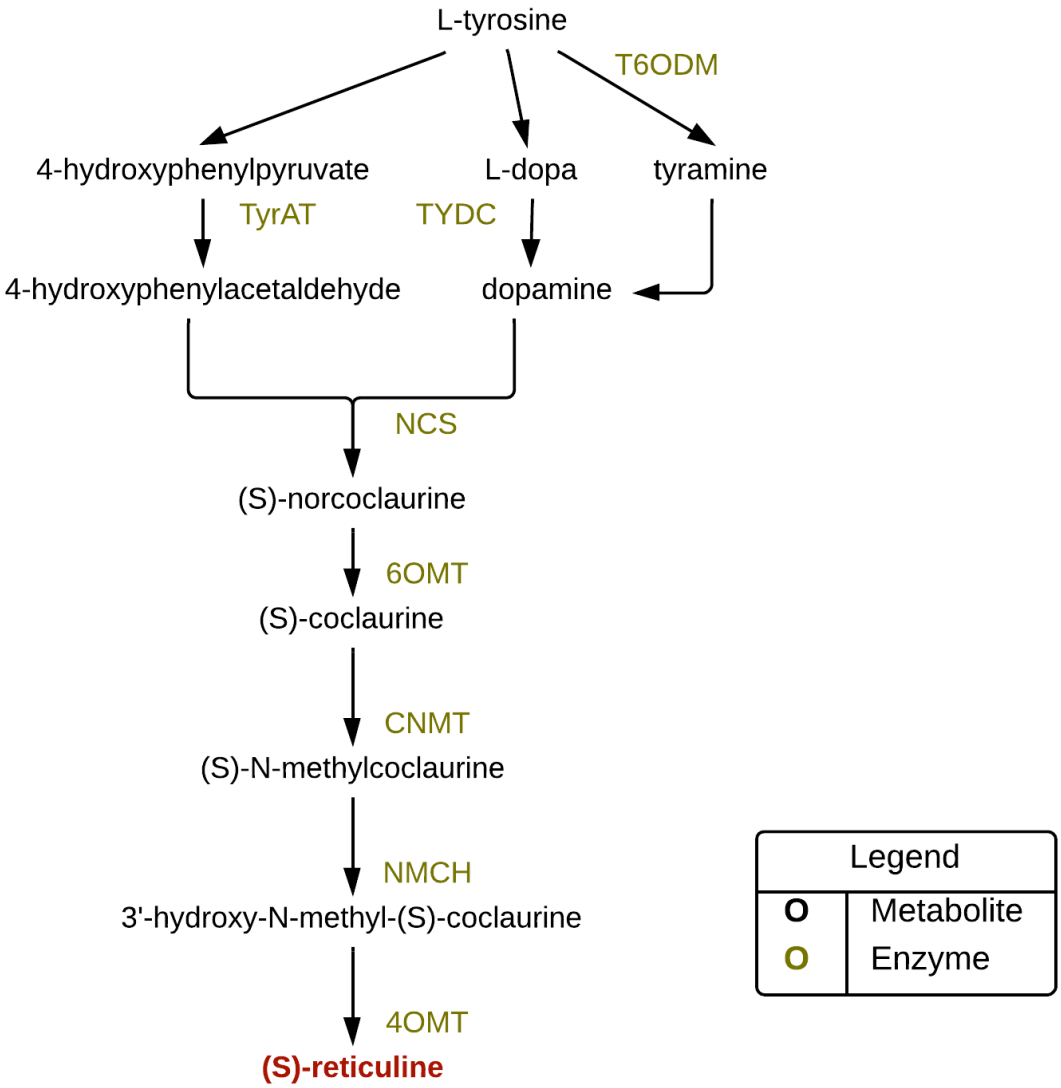


Figure 1.1: Graphical representation of the (S)-reticuline biosynthesis pathway
The biosynthetic pathway producing (S)-reticuline which is central to all benzylisoquinoline metabolites that are the focus for this study. Metabolites are featured in black and enzymes are featured in green. (S)-reticuline is highlighted red.

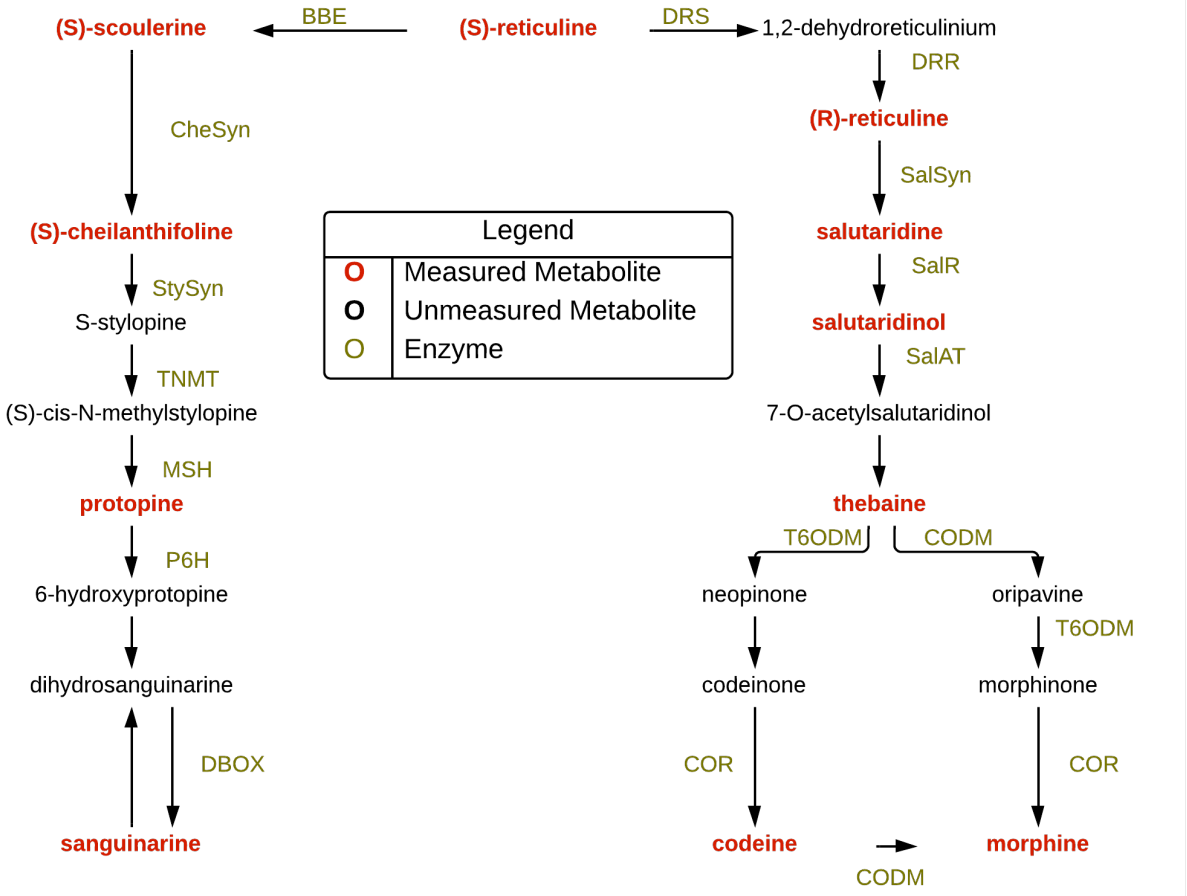


Figure 1.2: Graphical representation of the BIA biosynthesis pathway in *P. somniferum*
The benzylisoquinoline biosynthesis pathway starting with (S)-reticuline. Metabolites that had non-zero LC-MS/MS concentrations are featured in red. Metabolites colored in black were not analyzed for this study. Enzymes are featured in green.

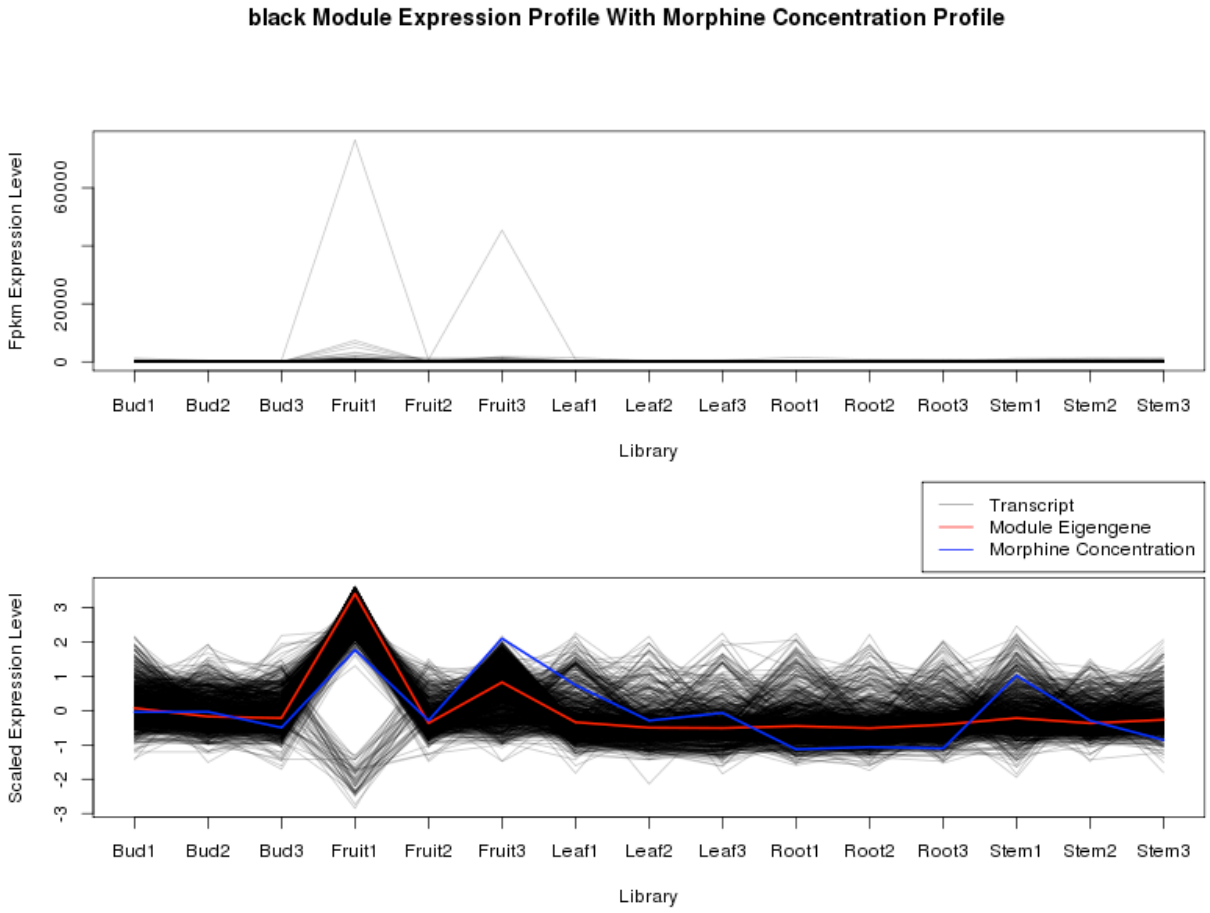


Figure 1.3: The relationship between the black module and the black module eigengene The upper plot shows the standard scale expression profile of all transcripts considered to be within the black module as identified by WGCNA. The lower plot shows the same transcription values standardized. The red line is the module eigengene which can be thought of as the first principle component of the expression profile of all genes in the black module. The blue line is the standardized morphine metabolite concentration.

CHAPTER 2

MATERIALS AND METHODS

In this study we intend to identify candidate transcripts that can be implicated as having a regulating effect on metabolite concentration. These transcripts may be isolated and identified as enzymatically involved in BIA biosynthesis or function as transcription factors that effect the regulation of our metabolites of interest. To this end we sample 5 tissues of Papaver somniferum across 3 biological replicates. cDNA libraries are built from isolated RNA and metabolites concentrations are measured. Sequencing and Assembly follow and finally the transcriptomic and metabolomics data sets are statistically analyzed for network structure, module membership, and module-metabolite relationships.

Plant Growth and Tissue Collection

Papaver somniferum plants were cultivated in the green houses in the Donald Danforth Plant Science center. The plants were subjected to an average temperature of 70°C and a maximum humidity of 40%. Supplemental lighting was provided at an intensity of 150 μMol that activated at any time in which the available sunlight dropped below 400 W/m^2. During the period of May 15th to September 15^{th} the supplemental lighting was active from 6:00 AM until 10:00 AM. During the remainder of the year the supplemental lights were active from 6:00 AM until 10:00 PM. The first batch of *Papaver somniferum* plants were seeded in spring. This batch had to be replanted in late summer due to an infestation.

Root, stem, leaf, flower bud, and developing fruit tissue samples were harvested over the winter. Tissue was harvested whenever it was present and the plant was in healthy condition.

Preferential sampling was done on new leaves and stems to encourage a broader range of transcriptional activity and larger RNA concentrations. Tissue samples were immediately frozen in liquid nitrogen and stored at -80°C until RNA or alkaloids were extracted.

RNA Sequencing and LC-MS/MS Alkaloid Measurement

RNA was extracted from tissue samples and sent to The University of Georgia. Replicate one samples were sent to the Beijing Genomics Institute (Hong Kong) for cdna library construction and sequencing using internal protocols. Replicate two and three isolated RNA samples were prepared for sequencing using an Illumina TruSeq kit. The resulting cDNA libraries were validated with the Aligent 2100 BioAnalyzer platform. Quantative PCR was performed to ensure proper pooling. Samples were pooled in 0.1% Tween solution. The samples were pooled into two lanes and sequenced in tandem.

Replicate one libraries were sequenced earlier than replicates two and three at the Beijing Genomics Institute (Hong Kong) on an Illumina Genome Analyzer machine resulting in a 75 base pair long paired-end read data set. Replicates 2 and 3 were processed at the Beijing Genomics Institute Americas (Davis, CA) and sequenced on a newer generation Illumina HiSeq 2000 machine. The resulting paired-end read data sets were stripped of barcodes and returned to the University of Georgia where they were stored on local databases until later analysis was performed. These datasets combined across replicates and tissues resulted in roughly two-hundred million paired-end reads.

The same tissue samples were ground under liquid nitrogen in order to extract alkaloids. The resulting powder was diluted with 70% ethanol and mixed. The solution was placed in a centrifuge and exposed to a 14,000 x g centrifugal force for 10 minutes at room temperature. The supernatant liquid was then filtered through a 0.2 μm low protein binding hydrophilic LCR

membrane. Liquid chromatography tandem mass spectrometry (LC-MS/MS) was performed on the isolated alkaloid samples. The peaks were normalized and Table 2.1 displays the resulting concentrations in μM.

Analysis Protocol

The analysis procedure for the *Papaver somniferum* RNASeq data is described in detail in figure 2.1. Reads for all replicate one libraries were in fastq-solexa format and all replicate two and three libraries were in fastq-sanger format. First all replicate one reads are convered to fastq-sanger format. Next reads are filtered using the FastX toolkit programs fastq-trimmer and fastq-filter. Fastq-trimmer removes low quality bases (quality score less than twenty) from the 3' end and reject reads that are trimmed below a minimum length of forty. Fastq-filter rejects reads that do not have a quality score of at least twenty over eighty percent of the read. The program Flexible Adapter Remover (FAR) is used to remove any remaining adapter sequence on the reads.

Assembly procedures are tested in two ways. First reads were processed by pooling and assembling them in tandem using the Trinity de novo assembler with default parameters (Grabherr et al., 2011). A second approach was to use the Trinity digital normalization pipeline to normalize the k-mer coverage of the assembly to 30 by rejecting reads with over represented k-mers. This has been shown to greatly reduce the running time for assembly and reduce the prevalence of sequencing errors in the final set of assembled sequences. Reads for the second assembly were separated into four files, two to contain all properly mated pairs and two to contain those reads, left or right orientation, that were orphaned during the quality control steps. Properly mated reads were digitally normalized which reduced the trinity input read file size by a

factor of ten. Orphaned reads were added to the digitally normalized reads and the reduced read set was assembled with the Trinity de novo assembler with default parameters.

Two options were explored for abundance estimation. For each assembly the original reads were mapped back to the assembled sequences and abundance estimation was done using RSEM (B. Li & Dewey, 2011). Alternatively, for each assembly an additional sequence filtering step was undertaken. Trinity assembled sequences were searched using the ESTScan program to reduce the number of assembled sequences that are falsely predicted (Iseli, Jongeneel, & Bucher, 1999). These potentially false sequences are removed by rejecting any sequence without a sufficiently large open reading frame. Sequences that fail ESTScan filtering were blasted against the NCBI non-redundant protein and nucleotide databases. Any sequence with a match having an e-score less than e-10 for protein matches or e-100 for nucleotide matches were reintroduced into the accepted sequence set. Finally RSEM was used to measure abundance values by mapping the original reads against these curated assembled sequence sets. Expectation maximized counts and length normalized counts (FPKM) were obtained for all four combinations of assembly and filtering procedure (B. Li & Dewey, 2011).

Figure 2.2 shows the effect that digital normalization and sequence filtering has on the assembled transcriptome size. The non-normalized Trinity assembly took roughly three weeks to finish execution and contained just under three-hundred and thirty thousand assembled sequences. This is in stark contrast to the roughly eighty thousand assembled sequences obtained by assembling only the first replicate tissue libraries in a previous study. The larger number of predicted sequences can be the result of many phenomena including sequencing error, false prediction of isoforms, or lack of sufficient coverage depth.

The sequence filtering protocol should correct for sequencing errors and false isoform prediction. Ideally this procedure should not remove biologically relevant sequences. If this were the case we would expect to see no change in the number of reads that are successfully mapped back to the contiguous sequence set by RSEM during the abundance estimation step. Figure 2.3 shows the effect that sequence filtering has on the percentage of successfully mapped paired-end reads. Surprisingly filtering has a positive effect on read mapping rates with an average read-mapping rate increase of 1.13%. Furthermore the difference in the percentage of reads mapped between the non-filtered and filtered contiguous sequence set is found to be statistically significant (matched pairs t-test with p=6.11e-08).

Comparison of the digitally normalized and non-normalized assemblies is difficult because there is no direct many-to-few relationship between the larger non-normalize assembly and the smaller normalized assembly. Digital normalization reduced the assembly size by nearly a half. If the reduction in assembly size is due to the removal of sequencing error driven falsely predicted sequences then we would expect to see an ubiquitous increase in expression level across the normalized assembly when compared to the non-normalized assembly. Figure 2.4 shows the relationship between the abundance measures of the digitally normalized and non-normalized assemblies. The average abundance measure is clearly greater in the digitally normalized assembly. The expression levels of transcripts represented in the normalized assembly are universally in greater abundance. This implies that even though nearly half of the assembled sequences are lost most reads are still mapped to a location in the transcriptome. This is considered evidence to support the notion that the lost sequences are most likely not biologically relevant.

Differential expression analysis was done at the gene level using the edgeR Bioconductor module (Robinson, McCarthy, & Smyth, 2010). Genes that did not have at least ten expected counts per million across all libraries and isoforms were immediately excluded from analysis. The EdgeR R package estimates the mean the expression level for a set of genes by estimating the mean expression level of each gene given its observed abundance values and a common parameter among all genes known as dispersion. Dispersion is defined as the squared coefficient of variation. EdgeR estimated a dispersion of roughly 0.194 between replicates which is interpreted as a 44% variability in abundance between replicated samples. A dispersion of 0.194 is large but fairly typical for RNASeq experiments on separate biological replicate organisms grown in a common environment. A large estimation of dispersion however does suggest that additional replication would benefit the analysis.

A test of no difference across all tissues was performed and all p-values were transformed to false discovery rates (FDR)(Benjamini & Hochberg, 1995). Isoform level transcripts belonging to genes with a FDR score less than 0.05 were passed on to co-expression network analysis with the WGCNA Bioconductor module. Differential expression filtering resulted in the elimination of 23,671 (58.84%) groups of transcripts at the gene level. This corresponds to the elimination of 62,961 (65.84%) transcripts at the isoform level and leaves 32,667 isoform level transcripts to continue into network formation.

Weighted co-expression network construction was done with a greatly reduced transcript set at the isoform level using length normalized FPKM abundance values. Automatic one-step network construction with the minimum module size set to 20 was run on the screened isoform transcript abundance profiles. The resulting modules were compared to the LC-MS/MS

generated alkaloid concentration levels to identify those modules that were significantly associated with secondary metabolite regulation.

Lastly the transcripts that were most highly associated with each module were scanned using the tool InterProScan (Zdobnov & Apweiler, 2001). InterProScan utilizes databases of known protein signatures in order to infer transcriptional function. Gene Ontology terms were gathered for most sequences. The Gene Ontology terms identified by InterProScan were used to develop GOSlim profiles for each module. These profiles are featured in Appendix A.

Table 2.1: LC-MS/MS alkaloid concentration values for *Papaver somniferum* samples

The concentration values are reported in μM. Any alkaloid with no observed concentration value may have been below the detection threshold for the LC-MS/MS procedure.

Tissue-sample	Sanguinarine	Macarpine	Protopine	Chelirubine	N-Methylstylopine	Scoulerine
leaf 1	0	0	0.01468	0	0	0
leaf 2	0	0	0	0	0	0
leaf 3	0	0	0	0	0	0
cap 1	0	0	0.00257	0	0	0
cap 2	0	0	0	0	0	0.00397
cap3	0	0	0	0	0	0.00167
bud 1	0	0	0.00406	0	0	0
bud 2	0	0	0	0	0	0.00083
bud 3	0	0	0	0	0	0.00167
stem 1	0	0	0.00307	0	0	0
stem 2	0	0	0	0	0	0.00667
stem 3	0	0	0	0	0	0.00067
root 1	0.73939	0	0.25705	0	0	0
root 2	1.44928	0	0.11673	0	0	0.00167
root 3	0.26812	0	0.07782	0	0	0.00333

Table 2.1 Continued

Tissue-sample	Cheilanthifoline	Stylopine	10-HDHS	Reticuline	Thebaine	Codeine	Morphine
leaf 1	0.00041	0	0	0.35357	16.11111	1.31622	6.01852
leaf 2	0.00047	0	0	0.01875	4.625	0.625	3
leaf 3	0.0013	0	0	0.02375	2.5	0.3125	3.65
cap 1	0.00035	0	0	0.44048	34.07407	1.94717	9.02778
cap 2	0.00099	0	0	0.155	7.125	0.83333	3
cap3	0.00119	0	0	0.2375	12.5	1.95833	10
bud 1	0.04229	0	0	0.07738	14.18519	0.87723	3.72917
bud 2	0.03896	0	0	0.1625	2.25	0.41667	3.75
bud 3	0.11039	0	0	0.1875	3.625	0.29167	2.4
stem 1	0.00027	0	0	0.36548	33.88889	1.44345	6.8287
stem 2	0	0	0	0.2375	9.625	0.83333	3
stem 3	0	0	0	0.04625	2.5	0.1125	1.35
root 1	0.00077	0	0	0.49643	0.64815	0.63021	0.55208
root 2	0.00182	0	0	0.0375	0.75	0.5625	0.75
root 3	0.00117	0	0	0.1875	0.5	0.41667	0.63

Table 2.1 continued

Tissue-sample	Salutaridine	Rhoeadine	Salutaridinol	Berberine
leaf 1	0.02469	0	0	0
leaf 2	0	0	0	0
leaf 3	0	0	0	0
cap 1	0.06272	0	0	0
cap 2	0.00833	0	0.00167	0
cap3	0.04167	0	0.005	0
bud 1	0.0121	0	0	0
bud 2	0.00694	0	0	0
bud 3	0.00417	0	0	0
stem 1	0.02778	0	0	0
stem 2	0.02778	0	0.00417	0
stem 3	0	0	0	0
root 1	0.00753	0	0	0
root 2	0.00694	0	0	0
root 3	0.01389	0	0	0

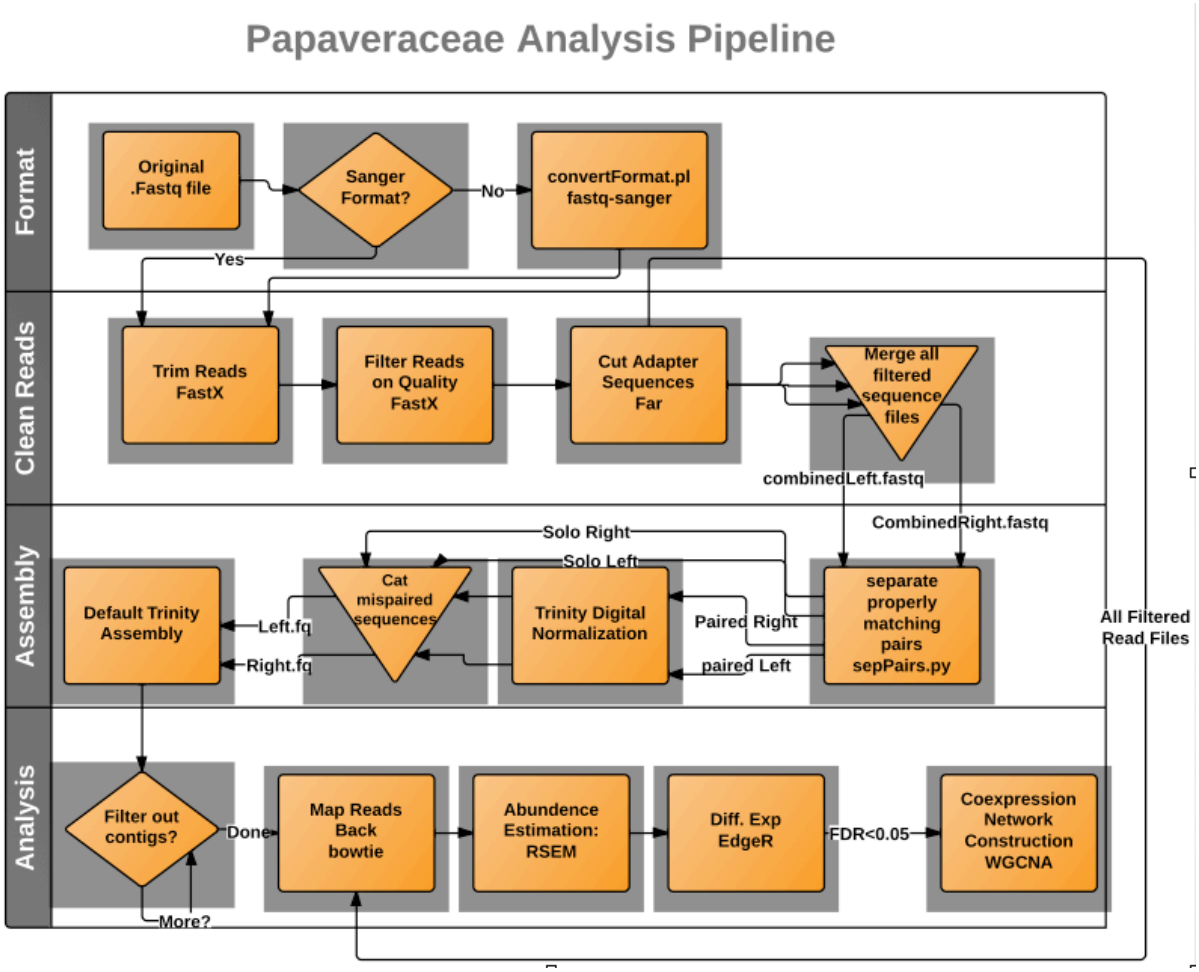


Figure 2.1: A schematic of the analysis protocol for P. somniferum sequencing data
A schematic diagram of the analysis protocol used to assemble and process *the P. somniferum* RNASeq data. The pipeline is broken into four main categories, Format, Cleaning Reads, Assembly, and Analysis. The Format step standardizes the input data type. The Clean Reads step removes data of low quality. The Assembly step assembles all remaining reads into a complete pooled transcriptome. Finally the analysis step is comprised of abundance estimation, differential expression analysis, and finally co-expression network construction.

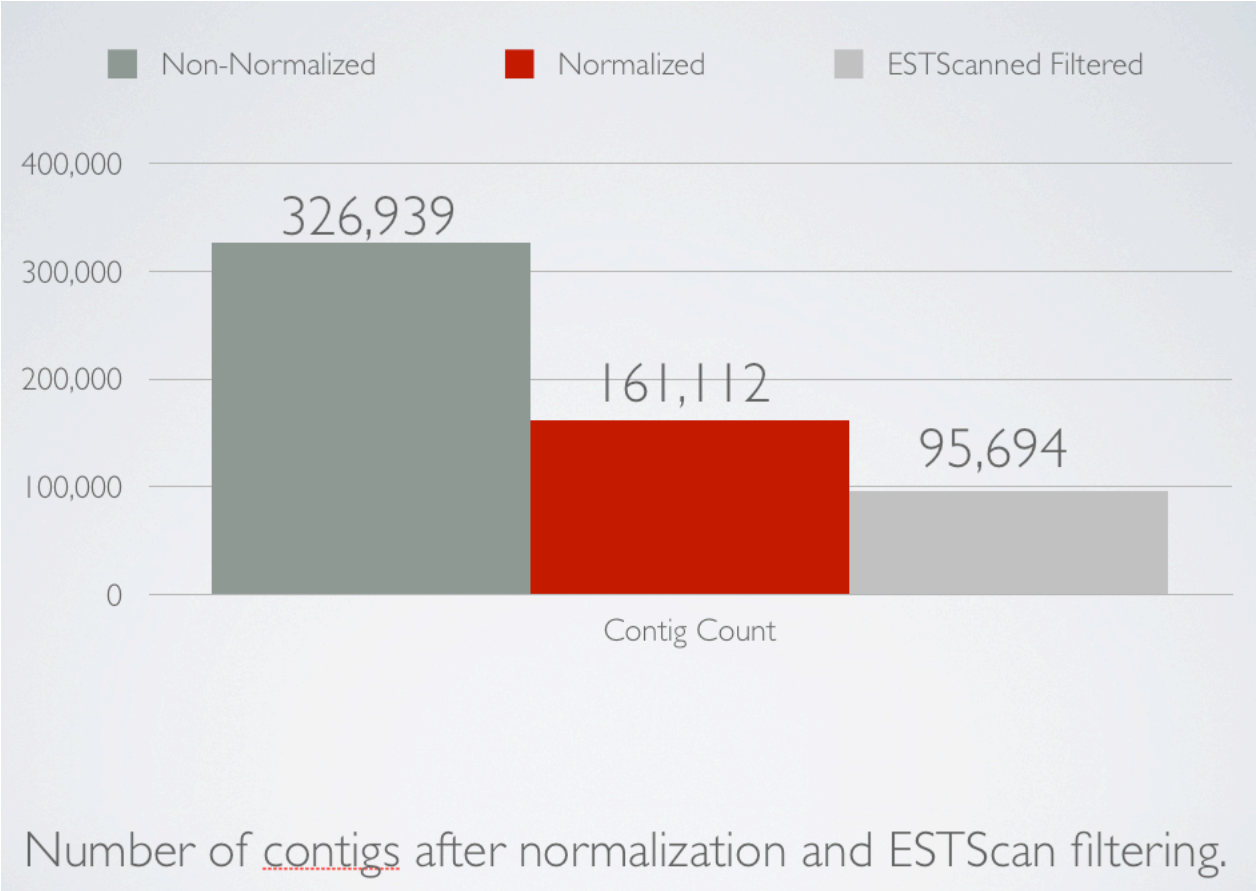


Figure 2.2: The effect of filtering strategies on transcriptome size
The dark grey bar shows the number of sequences assembled in the non-normalized and non-filtered Trinity assembly. The number of contiguous sequences assembled after the application of digital normalization to the RNASeq reads is shown in red. The light-grey bar shows the number of sequences of the remaining normalized sequences that pass the ESTScan filtering step

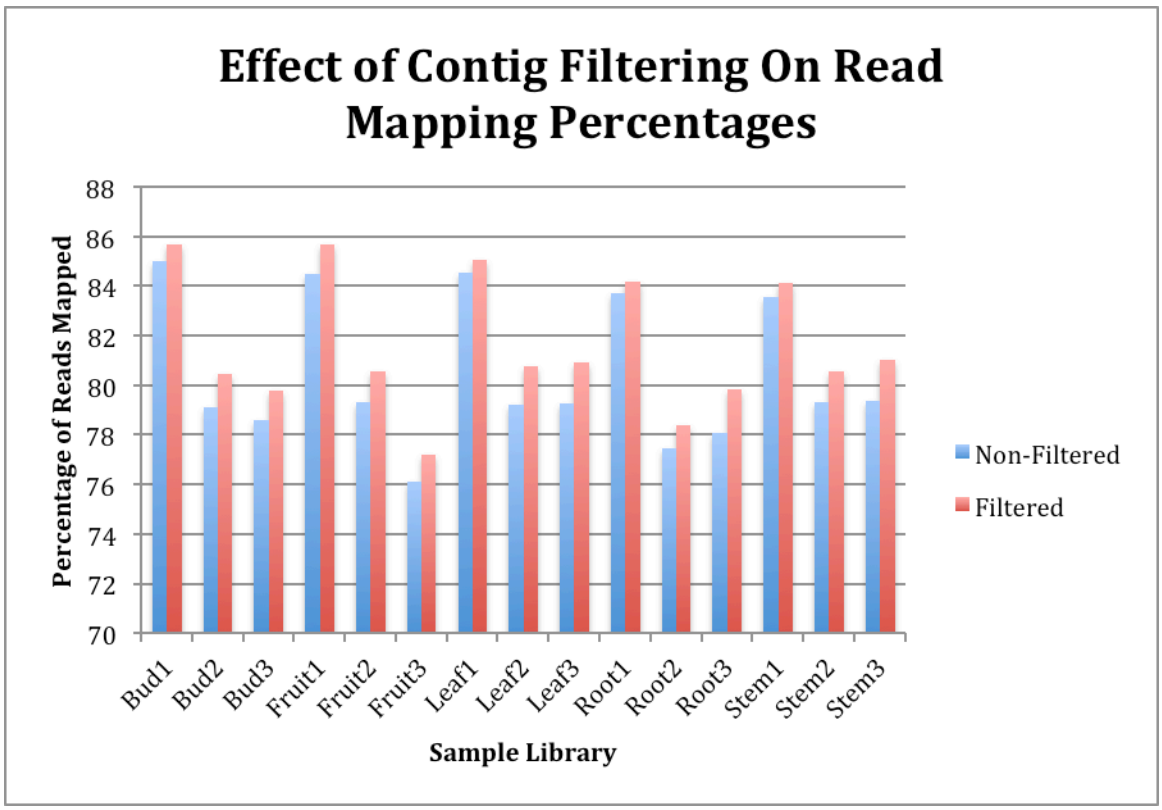


Figure 2.3: Effect of Contig Filtering On Read Mapping Percentages
Percentage of reads mapped back to the trinity generated contiguous transcript sequences (contigs) during the RSEM abundance estimation step for non-filtered and filtered assemblies. Filtering has a uniformly positive effect on the read-mapping rate across all sample libraries.

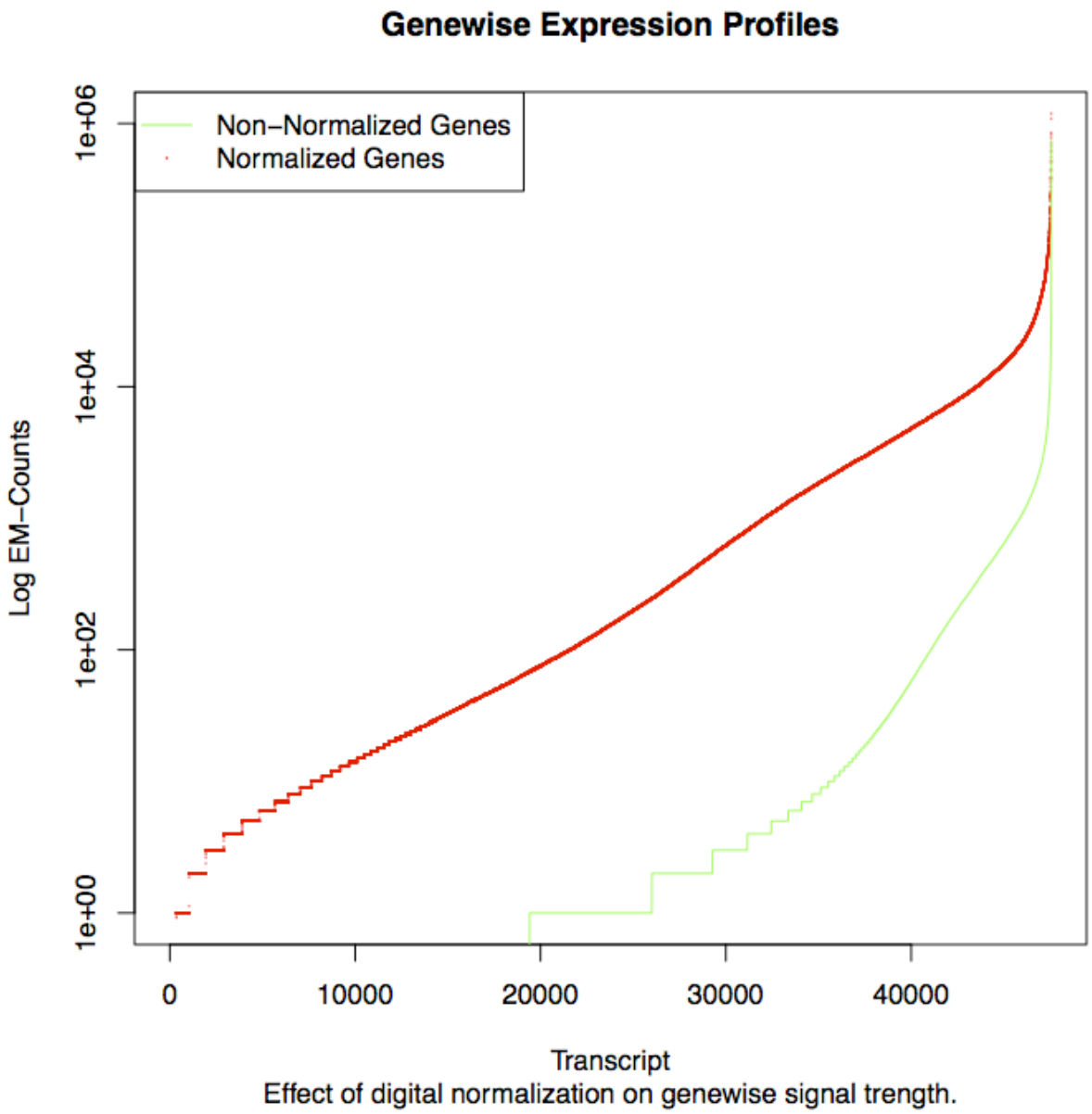


Figure 2.4: The abundance levels of the normalized and non-normalized assemblies
The green line is the gene-level ordered abundance values for all components in the non-normalized assembly. The red is the gene-level ordered abundances values for all components in the normalized assembly. It is important to note that the two lines are not on similar x-axis scales since there are roughly half as many assembled sequences in the normalized assembly.

CHAPTER 3

RESULTS AND DISCUSSION

Sample Clustering

Replicate libraries (samples) for all studied organs (roots, stems, leaves, flower buds and fruits) were clustered based on transcript abundance estimations. Figure 3.1 shows a multi-dimensional scaling (MDS) plot of the samples based on the 10,000 transcripts exhibiting the greatest degree of among-level variation in expression levels based on non-length normalized raw counts. MDS plots are analogous to principal component analyses and the two axes represent the two most influential principal components. Variation was observed among biological replicates for the same tissue. EdgeR detected an estimated 40% variance between replicates but in most cases tissue clusters show to be the most dominant factors. There is an interesting pattern of clustering between the fruit and stem tissues. Fruit replicate 2 appears to be much more similar to the stem tissues than the other fruit tissues.

Figure 3.2 shows a sample clustering of just those transcripts that were declared differentially expressed based on length normalized FPKM abundance measures (fragments per kilobase of sequence per million reads mapped). Again, fruit 2 clusters with stem and the fruit1 and fruit3 samples appear quite distant from the other samples. The lower panel of figure 3.2 shows the distribution of metabolite concentrations across the tissues. The strong difference in the first and third fruit samples suggests that there is a phenotypic difference between these samples corresponding to the differences observed in transcript abundances. In general, variation among tissues dominates clustering patterns based on transcript abundances. The corresponding distribution of alkaloid concentrations imparts confidence that the sampling strategy used for this

study should be effective in identifying transcript clusters with abundances that are correlated with benzylisoquineoline alkaloids (BIAs) and transcripts encoding enzymes contributing to their biosynthesis.

Transcript Clustering and Network Construction

More generally we are most interested in clusters of transcripts with similar expression patterns. Transcripts were initially hierarchically clustered based on the absolute value of the correlations between them as a distance measure. Modules were identified by partitioning the dendrogram that results from this clustering. This process is graphically represented in figure 3.3. Thirty-five modules were identified across the set of differentially expressed transcripts with a minimum module size of forty-seven transcripts, mean module size of eight-hundred and thirty-four transcripts, and maximum module size of three-thousand seven-hundred and thirty-two transcripts.

After modules were identified, a module eigengene was calculated for each module. Figure 1.3 illustrates the relationship between the expression profiles of constituent transcripts and the central tendency of the module summarized as an eigenvector, or eigengene. Figure 3.4 shows the significant ($p < 0.05$) relationships between the eigenvector for each module and the measure BIA concentrations. The use of eigengenes in order to calculate module-metabolite relationships is invaluable because it provides a summary of the dominant expression profile for all of the transcripts within a module (figure 3.9). The salmon, greenyellow, and skyblue modules have a similar relationship to all of the measured BIAs. Modules need not represent one single underlying biological process (e.g. BIA biosynthesis) and biological processes are likely not wholly contained in a single module. Many modules exhibit significant correlations between

eigengene vectors and BIA concentrations. Investigation of these models should provide insights into the regulation of BIA biosynthesis.

Known BIA-related enzyme coding genes

Most of the enzyme coding genes for the morphinan branch of the BIA biosynthesis pathway have been identified and were listed in chapter one of this text. One-hundred and two sequences were identified as having a significant nucleotide BLAST e-value when compared to the known nucleotide sequences for salutaridine synthase (SalSyn), salutaridinol reducatase (SalR), salutaridinol acetyltransferase (SalAT), thebaine 6-O-demethylase (T6ODM), codeine demethylase (CODM), and codeinone reductase (COR) (Altschul, Gish, Miller, Myers, & Lipman, 1990). Table 3.1 shows transcript assemblies with nearly identical alignments with the known morphinan branch enzyme coding genes.

Figure 3.6 shows the distribution of the sequences that are nearly identical matches among the modules. It is interesting that these sets of modules do not overlap. Figure 3.6 also shows three modules that exhibit the highest correlations between their eigengene expression values and BIA levels. Surprisingly, the T6ODM transcript did not show differential expression among tissue groups. Therefore, T6ODM was not included in network construction analyses. T6ODM seems to be ubiquitously expressed and may serve functions outside of morphine biosynthesis.

It is thought that the morphinans are produced throughout the plant and the secondary metabolites are not moved long distances (Toni Kutchan, personal communication, April 12, 2013). The implications of the dissonance between the known enzyme-containing modules and the modules most highly associated with the morphinan BIA concentrations can be numerous. In

order to explore a possible reason for this counterintuitive module to metabolite relationship for modules containing known morphinan biosynthesis enzyme coding we will examine the brown module. The brown module contains the transcript with a high identity match to the canonical CODM enzyme coding gene.

Brown module analysis

The brown module is the second largest module with 3027 constituent transcripts including codeine O-demethylase (CODM). Surprisingly, the brown module eigengene is negatively correlated with morphine with a Pearson correlation coefficient of -0.54 ($p=0.04$). This is counter intuitive given that this module contains an enzyme coding gene that is critical to the biosynthesis of morphine (figure 3.5). One could speculate on many biological reasons for this negative association but the simplest explanation may lie in the relationship between the module eigengene and the transcripts from which it was calculated in large modules – i.e. the CODM expression profile may not be well represented by the eigengene for the brown module.

The WGCNA analysis pipeline attempts to split modules into more tightly associated sub-modules and also merges modules that aren't significantly different from each other. This study is based on the correlation analysis of fifteen total samples (5 tissues X 3 biological replicates) which is a relatively small number given the large number of transcripts for which we are attempting to estimate correlations and the fairly high level of variation among biological replicates. A direct effect of the small number of samples is that only extreme positive or negative correlations become statistically significant. There is a clear subset of the transcripts contained in the brown module that track the up and down regulation of morphine more closely than the eigengenes for the three modules with the strongest morphine correlations. Furthermore the brown module eigengene is essentially the first principal component of the expression

patterns in that module. The second principal component may be nearly as strong as the first especially in modules of such large size. The bifurcated pattern (positive and negative correlations) around the stem libraries illustrates how some transcript abundance may be positively or negatively correlated with the eigengene profile. Additional sampling would allow separation of positive and negative correlations and finer clustering of genes into modules during network construction.

Figures 3.6, 3.7, and 3.8 detail the percentage module composition for a plant specific subset of gene ontology terms for the molecular function, biological process, and cellular component categories respectively for each module of interest shown in figure 3.6 and the overall distribution of the selected gene ontology terms. Tables A.1, A.2, and A.3 show the statistical significance of the up and down enrichment of these terms relative to the whole distribution of transcripts in the network. The brown module is most highly enriched for increases in protein modification processes, signal transduction, and nuclear activity. There are highly significant reductions in translational activity, photosynthesis related processes, structural molecular activity and ribosomal activity.

The brown module does include transcription factors and some of these may be involved in regulation of CODM. Recently the psWRKY gene was identified as a wounding responsive transcription factor controlling BIA biosynthesis (Mishra et al., 2013). Eleven transcripts with strong BLASTX e-value matches against psWRKY or related transcription factors were found in the brown module. The CODM transcript itself is highly correlated (PCC=0.84) with a transcript of strong similarity to the JOINTLESS gene in tomato with is a MADS-BOX type gene controlling tomato flower abscission zone development (Mao et al., 2000). The gene ontology term enrichment in signal transduction and protein modificationin the brown module, the

presence of many high quality matches for WRKY type transcription factors, and the inclusion of the only exceptional quality match for CODM each indicate the potential importance of the brown module in BIA biosynthesis. W-box WRKY target sites have been identified in the promoter regions of upstream BIA biosynthesis genes (4OMT, 7OMT, SAT, BBE and TYDC; Mishra et al. 2013). Future experimental research will assess the impact of silencing expression of these transcripton factors on BIA and specifically morphine biosynthsis. Efforts are under way to isolate and sequence the promoter region of the CDOM gene in order to characterize target sites for specific transcription factor families.

Conclusion

There is ample evidence to show that the clusters that have been identified have some underlying common biological function based on the levels of gene ontology term enrichment found across modules. The brown module has been implicated in some level of involvement in morphinan BIA biosynthesis by membership of the CODM enzyme coding gene and a collection of WRKY type transcription factors that have recently been shown to regulate BIA biosynthesis in response to wounding. Every module has at least one significantly enriched gene ontology term in the biological process and molecular function categories and all but three show significant enrichments for cellular component association. These widespread enrichments across modules is evidence that there is a real biological meaning underlying the transcript modules identified in the co-expression network. Given this validation of the biological relevancy of the *P. somniferum* weighted gene co-expression network we can then query the network for interesting novel gene associations.

Seven hundred and nine transcripts were found to have a strong nucleotide to protein BLAST similarity to a known or putative transcription factor in the NCBI non-redundant protein

database. There are possible transcription factors that clearly associate with every module within the network. Some of these potential transcription factors are strongly connected within modules and others are shared among modules that have similar relationships to metabolite profiles. Transcription factors found to co-express with other BIA biosynthesis enzyme coding genes are obvious targets for further study.

Similar screening can be done for many other gene classifications of interest and provides a statistically reliable filter of the large and complex transcriptome space of *P. somniferum*. This work may serve as a foundation for molecular evolutionary analyses within the Papaveraceae and across all plants. Gene screening, silencing, and overexpression analyses can be structured on this network in order to elucidate the finer interactions between targeted transcript expression and BIA concentrations. The function of potential homologs of the known enzyme coding genes is of particular interest and may be explored in the future. Most importantly though we have shown a proof of concept of the potential to construct strongly connected and biologically relevant gene co-expression networks in plants using carefully controlled and normalized de novo transcriptome assemblies of RNASeq data.

Table 3.1: BLAST hits to known enzyme coding genes
The sequences with nucleotide BLAST e-value scores of zero against the known enzyme coding genes in the *P. somniferum* morphinan biosynthesis pathway.

Enzyme	Contig	BLAST Alignment Length	BLAST e-value	BLASTN Percent Identical	Module
CODM	comp26126_c0_seq1	1427	0.00E+00	99.93	brown
SalSyn	comp55061_c0_seq5	1711	0.00E+00	99.42	cyan
salRed	comp47409_c0_seq1	1258	0.00E+00	99.92	cyan
salAT	comp49748_c1_seq5	1148	0.00E+00	99.04	darkred
salAT	comp49748_c1_seq3	818	0.00E+00	99.27	darkred
cor	comp56696_c0_seq35	565	0.00E+00	98.58	midnightblue
cor	comp56696_c0_seq30	565	0.00E+00	98.58	steelblue
cor	comp56696_c0_seq22	565	0.00E+00	98.58	turquoise
T6ODM	comp51790_c1_seq5	1229	0.00E+00	98.62	
cor	comp56696_c0_seq34	573	0.00E+00	97.38	
cor	comp56696_c0_seq24	573	0.00E+00	97.38	
cor	comp56696_c0_seq5	573	0.00E+00	97.38	

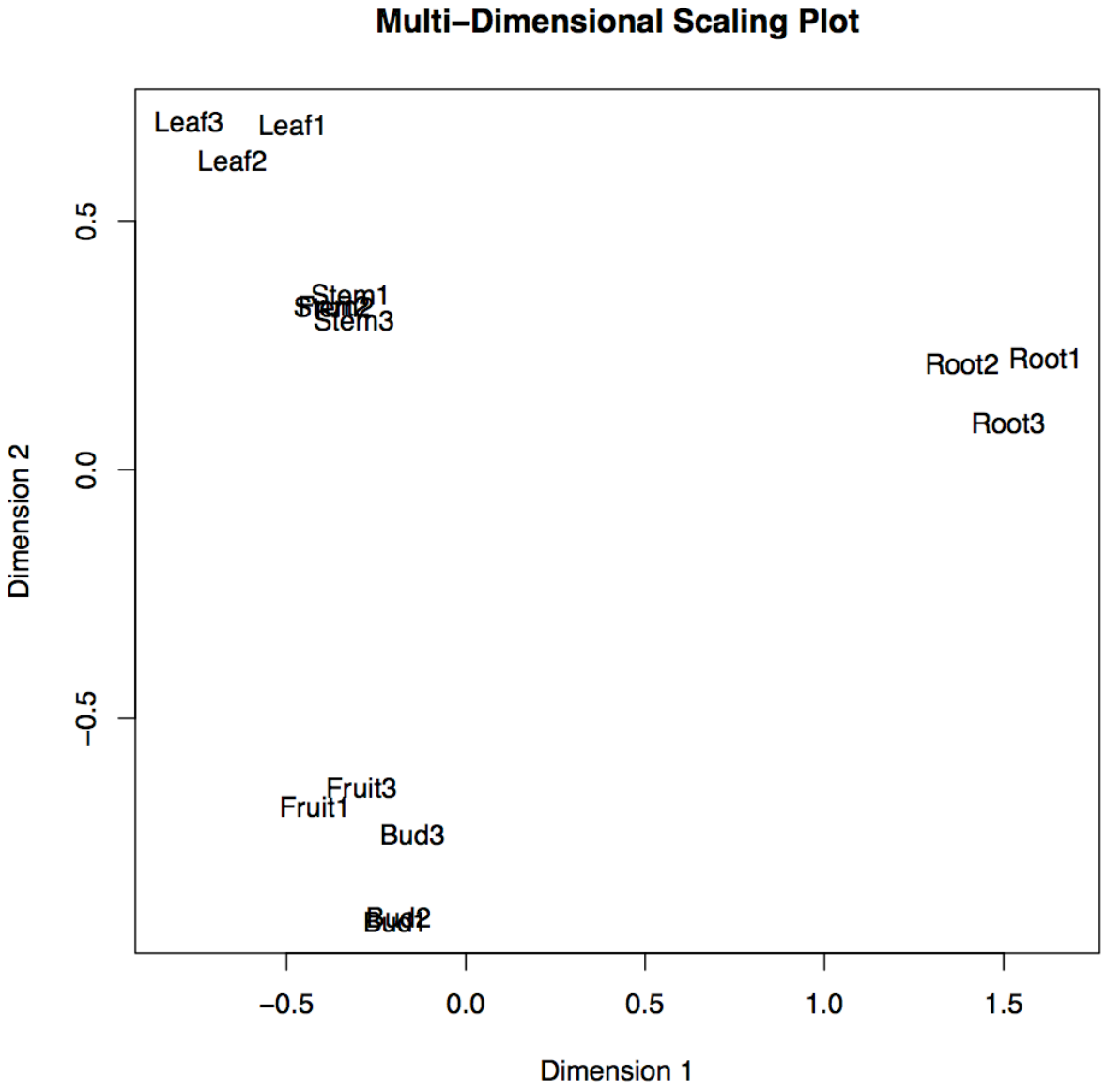


Figure 3.1: Multidimensional scaling plot of *P. somniferum* libraries
Distance based clustering based on the first two principal components of expression variation. The clustering based on tissue type suggests that differences between replicates do not need to be correct

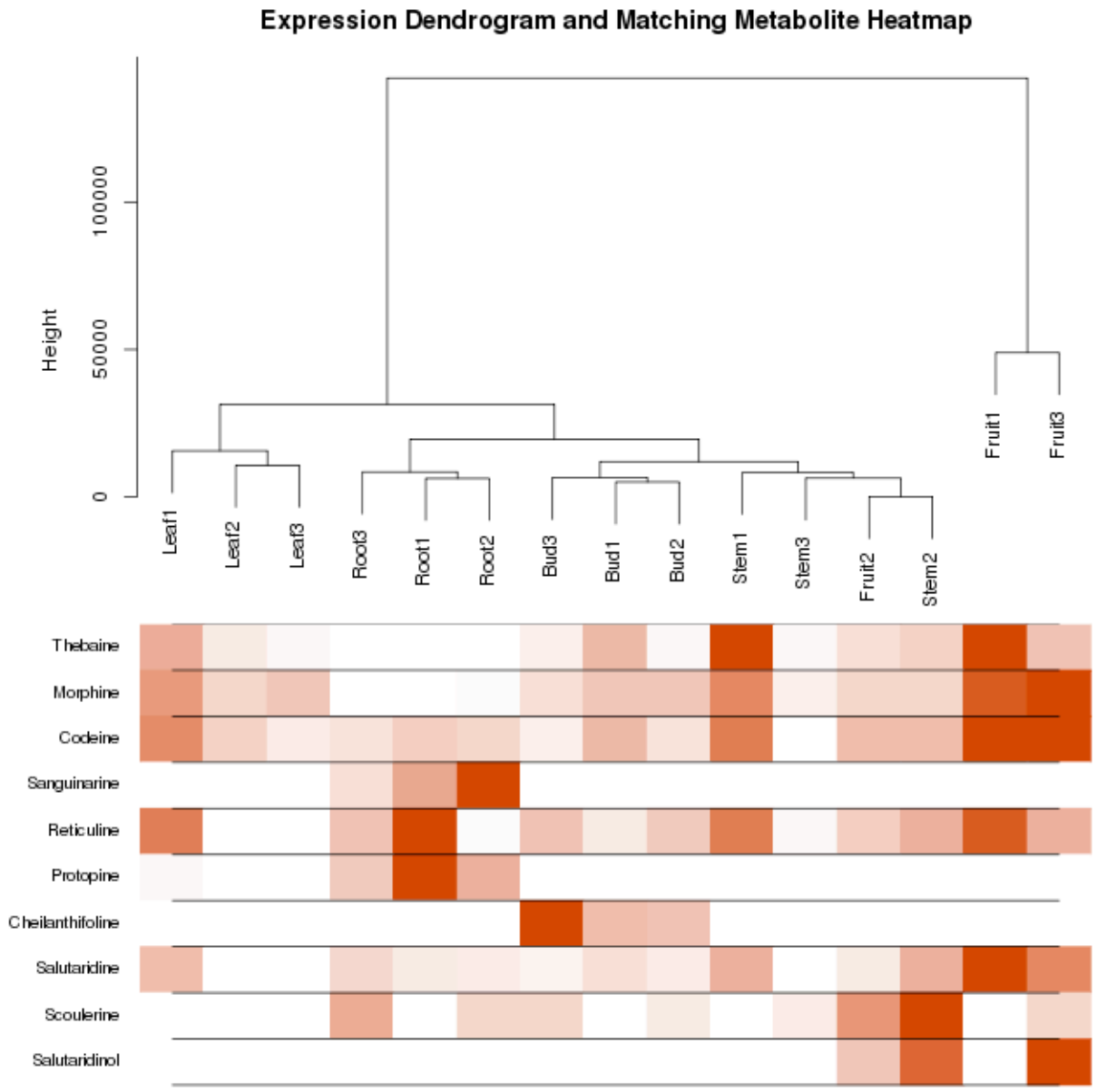


Figure 3.2: Expression dendrogram and matching metabolite heatmap
The dendrogram is based on the Euclidean distances between the fpkm length normalized abundance values for every transcript within each library. The metabolite concentration heatmap below is reordered to fit the dendrogram and is not considered when structuring the tree

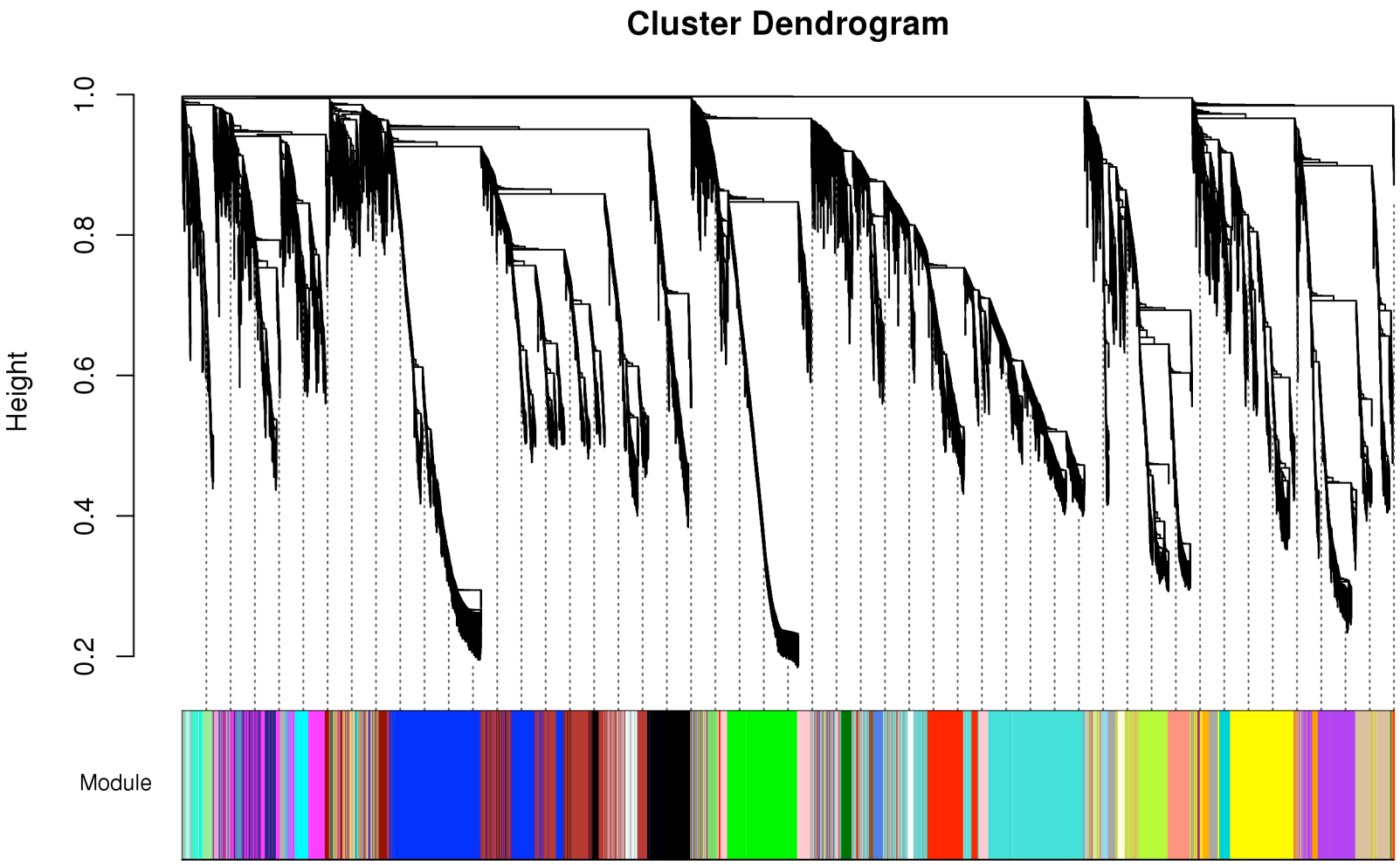


Figure 3.3: Transcript clustering and module assignment
Transcripts are clustered based on abundance and then grouped using the dynamic height cut method. Top: A dendrogram of transcripts based on expression. Bottom: Colors of the modules that each transcript was assigned to.

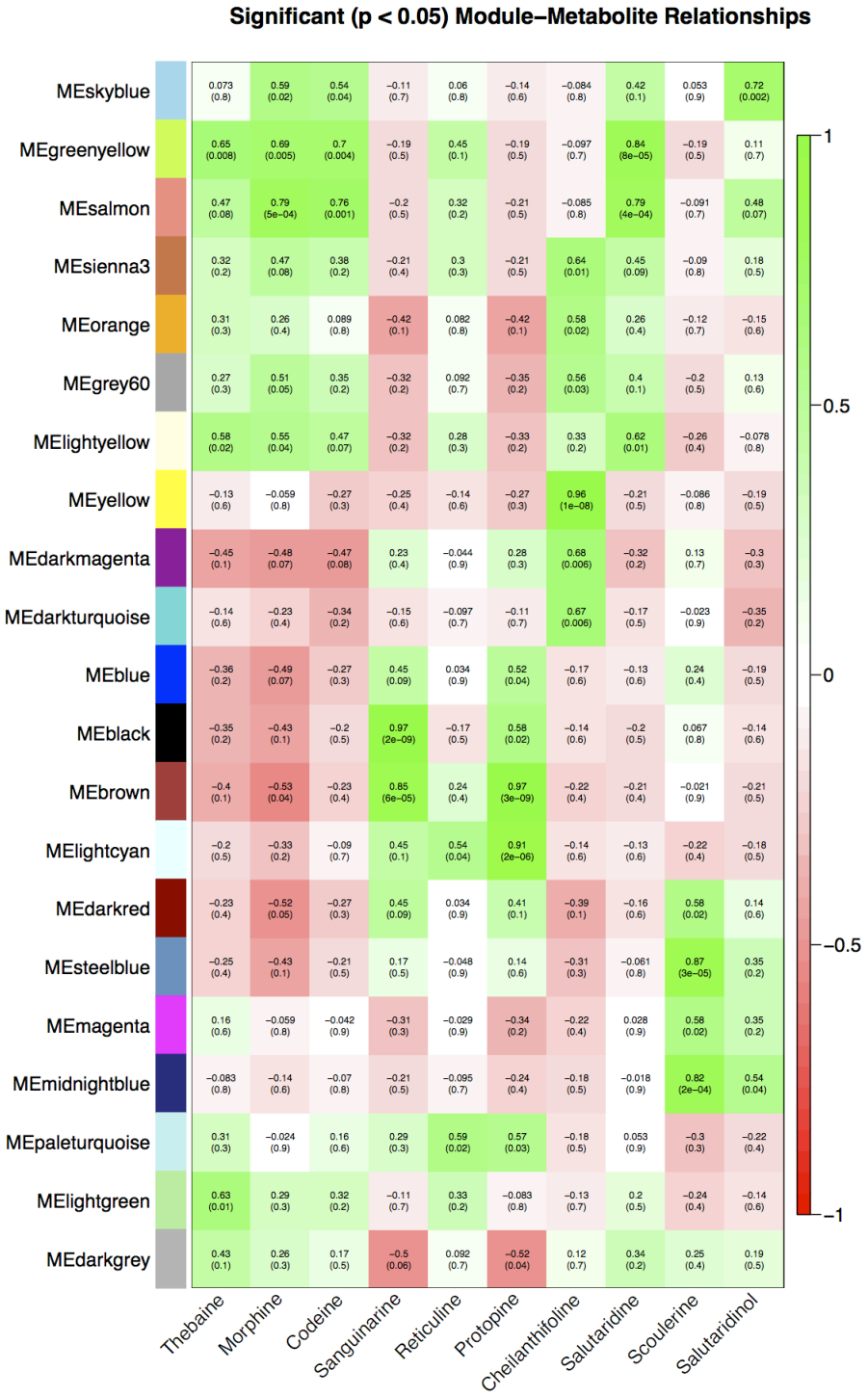


Figure 3.4: Significant module-metabolite relationships
A heatmap of the statistically significant correlations between each module eigengene and each measured concentration of morphinan BIAs. The green to red gradient represents strong positive to strong negative correlations.

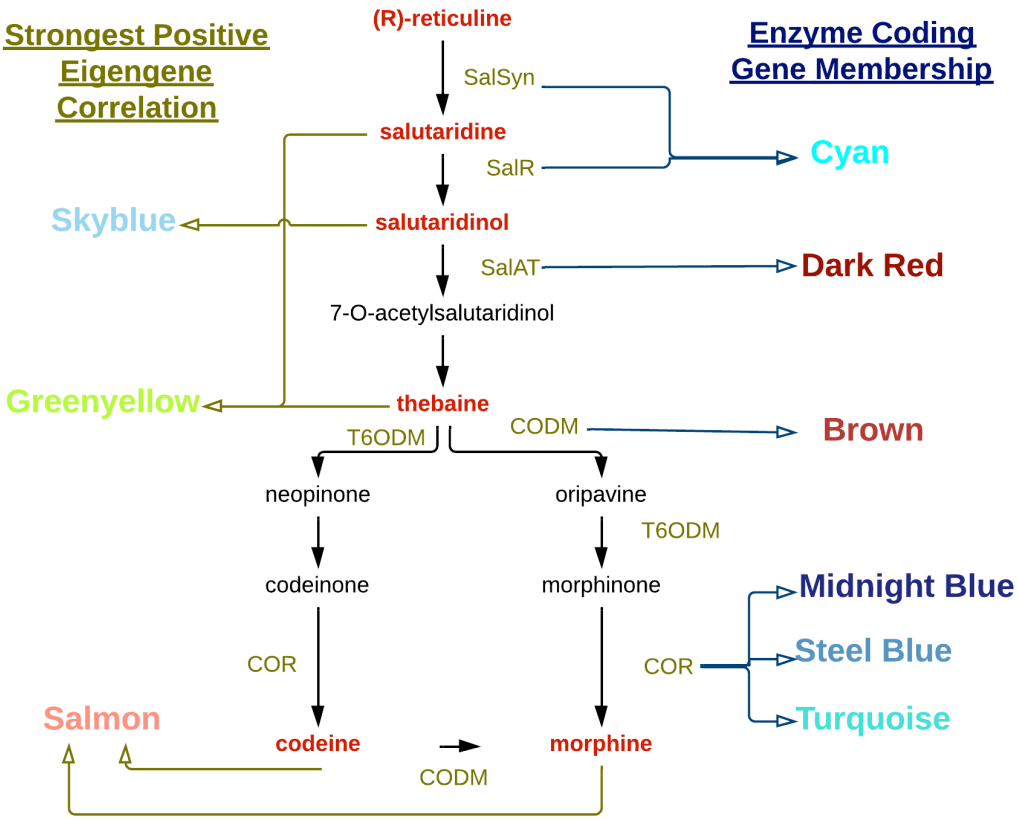


Figure 3.5: Distribution of morphinan enzyme coding genes across modules
A schematic diagram of the morphinan BIA biosynthesis pathway branch with module identifiers shown. Left: The modules with the most significant positive correlation with the measured morphinan BIA concentrations. Right: The modules containing transcripts with nucleotide BLAST hits with an e-value arbitrarily close to zero and hit identify greater than 98%.

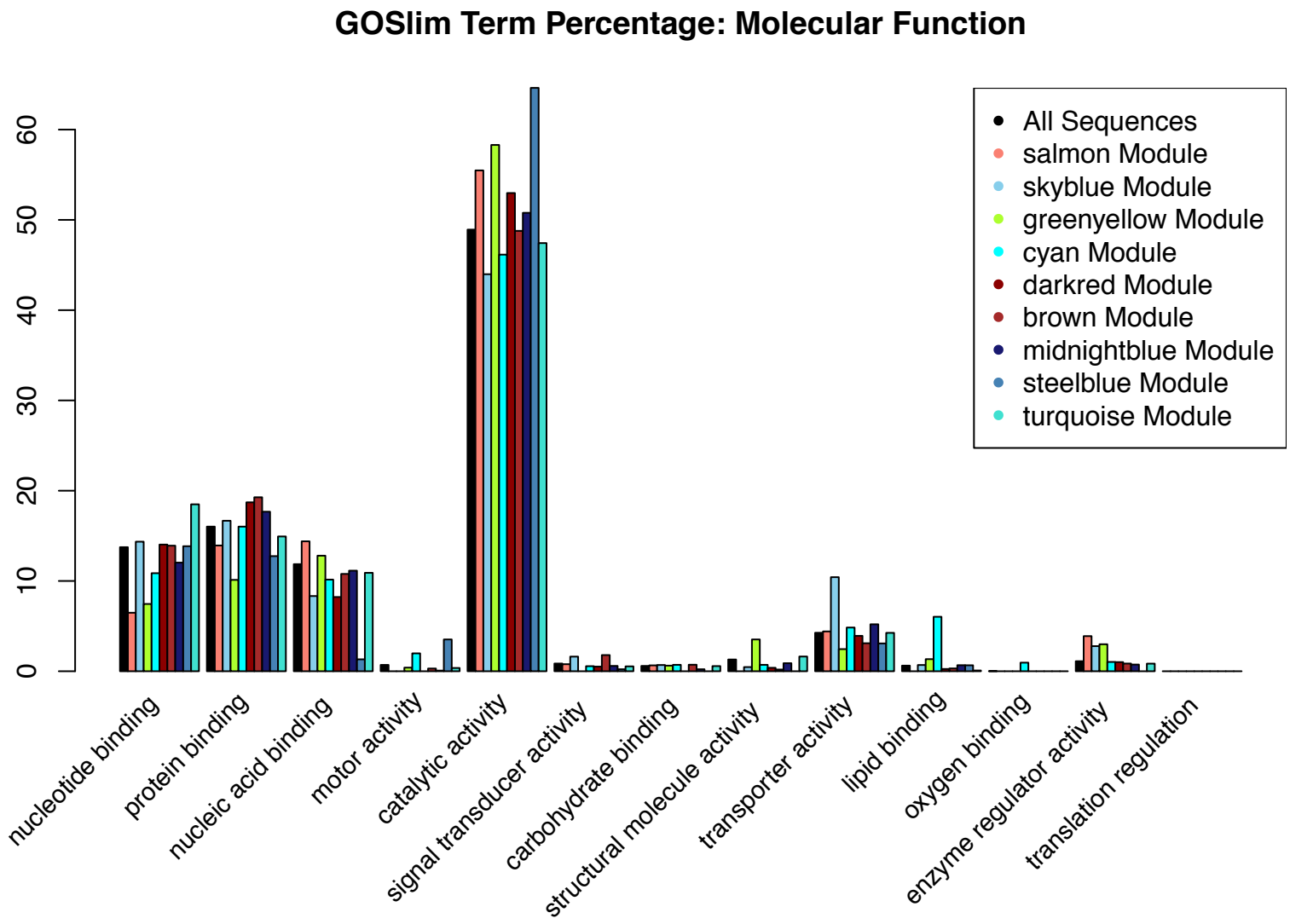


Figure 3.6: GOSlim term percentages for selected modules in the molecular function category
A plot of the relative GOSlim molecular function term composition for each module of interest by percentage.

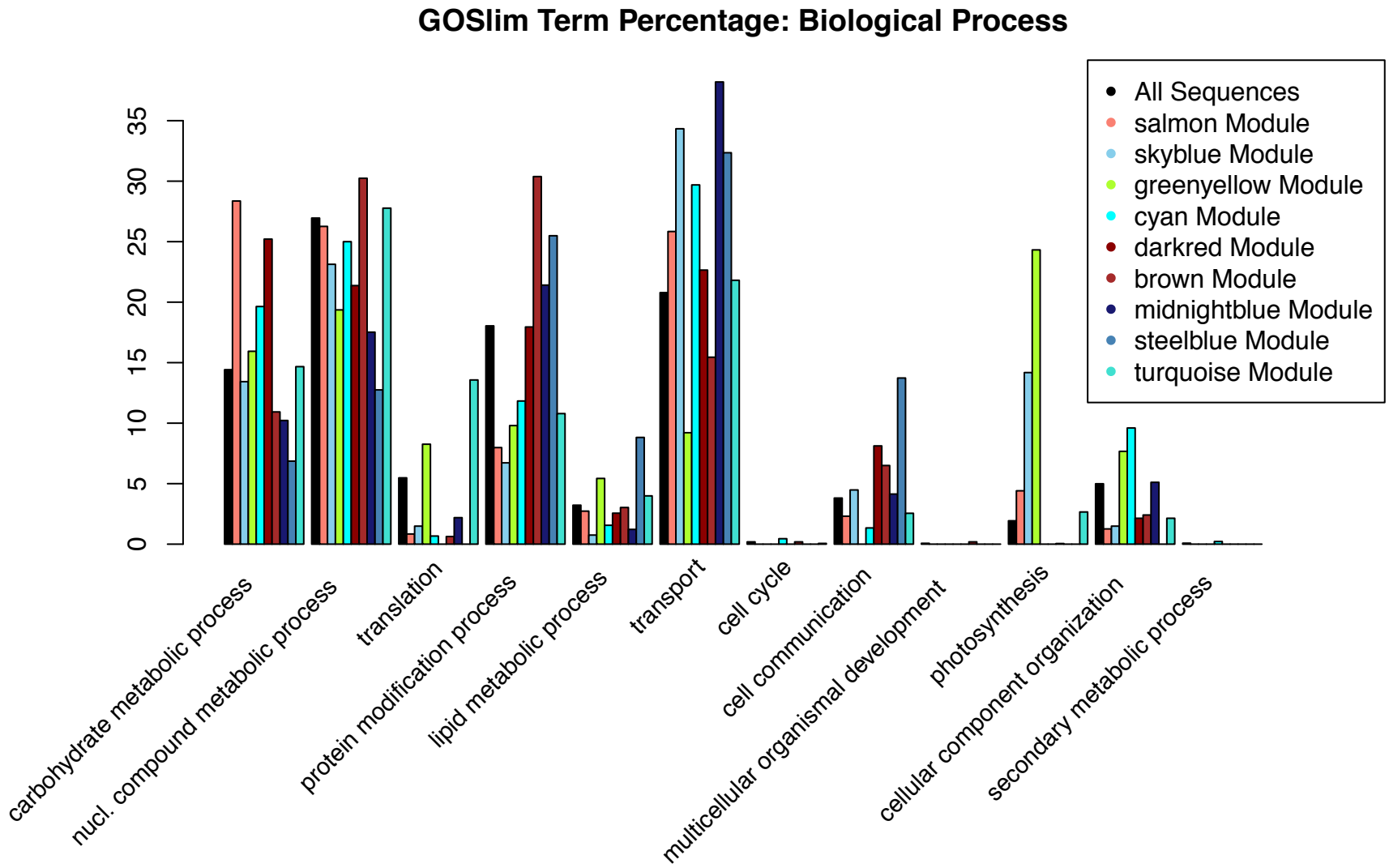


Figure 3.7: GOSlim term percentages for selected modules in the biological process category
A plot of the relative GOSlim biological process term composition for each module of interest by percentage.

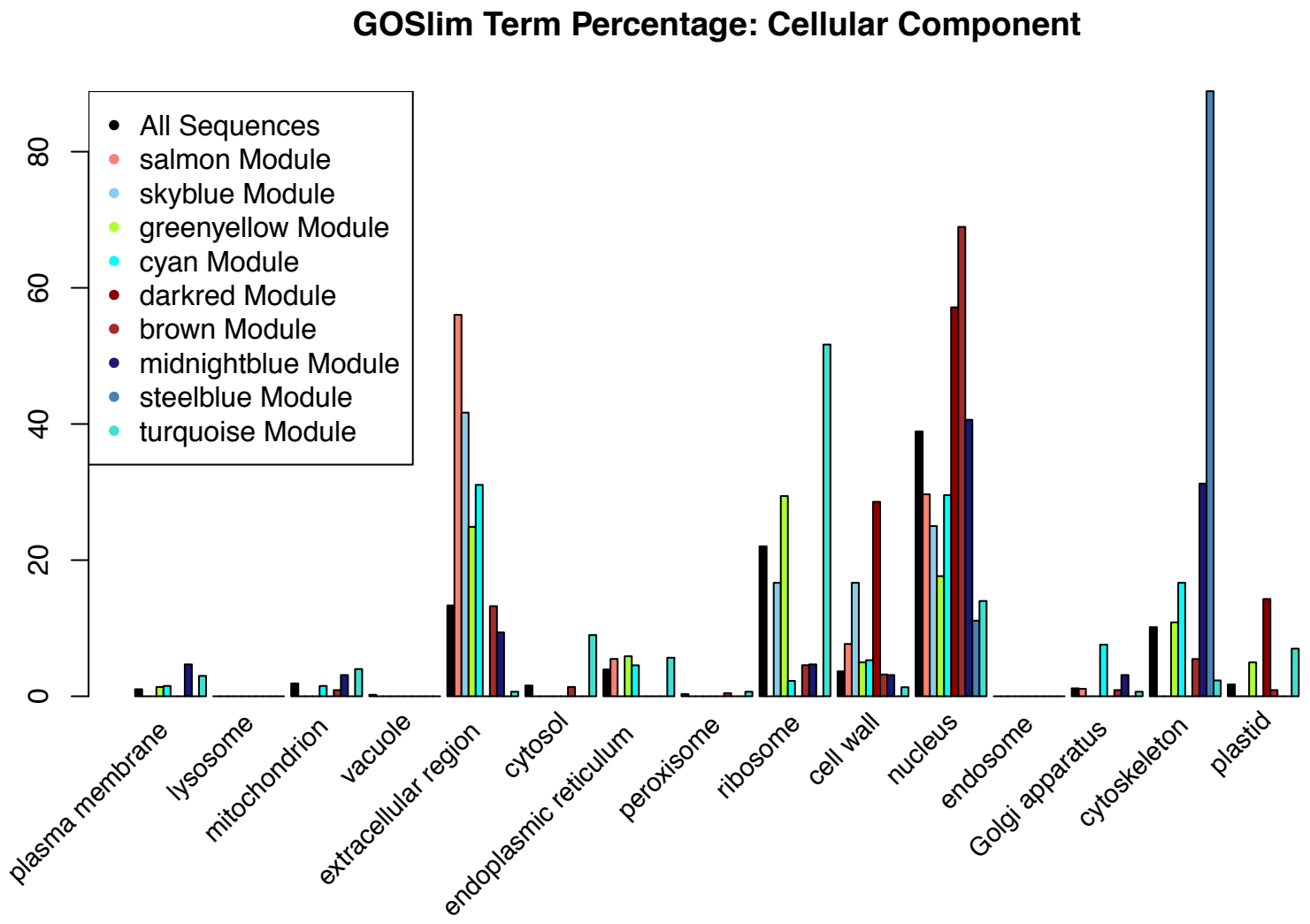


Figure 3.8: GOSlim term percentages for selected modules in the cellular component category
A plot of the relative GOSlim cellular component term composition for each module of interest by percentage.

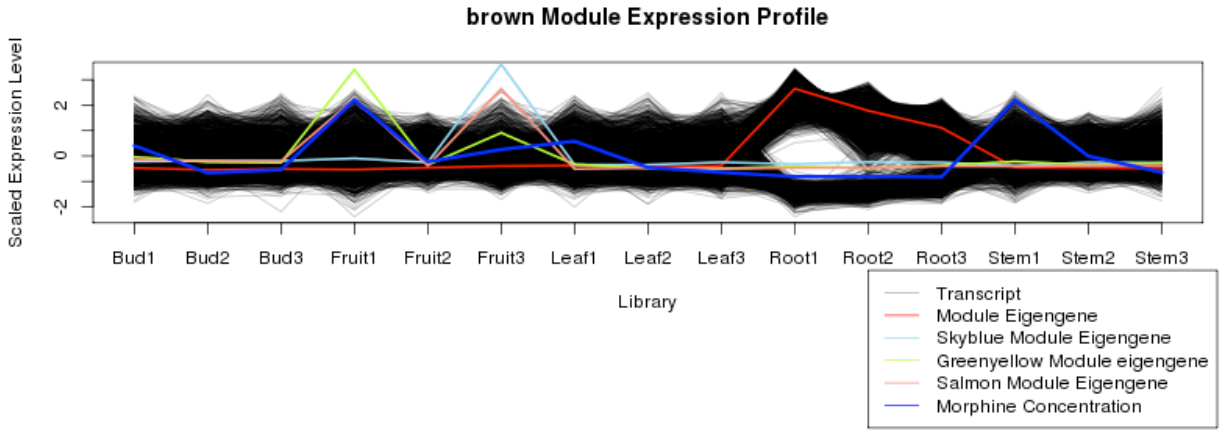


Figure 3.9: Transcript abundance profiles for the brown module with top morphine modules

A transcript abundance profile for all transcripts in the brown module with the module eigengenes for the tree top morphine correlated modules. Morphine concentration is shown in dark blue. Values for each transcript, eigengene, and concentration have been standardized in order to view relative up and down regulation on the same scale.

REFERENCES

Ahmad, N., Gupta, S., Husain, M. M., Heiskanen, K. M., & Mukhtar, H. (2000). Differential Antiproliferative and Apoptotic Response of Sanguinarine for Cancer Cells versus Normal Cells. *Clinical Cancer Research, 6*(4), 1524-1528.

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., . . . Giovannoni, J. J. (2005). Transcriptome and Selected Metabolite Analyses Reveal Multiple Points of Ethylene Control during Tomato Fruit Development. *The Plant Cell Online, 17*(11), 2954-2965. doi: 10.1105/tpc.105.036053

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Bauer, W., & H. Zenk, M. (1991). Two methylenedioxy bridge forming cytochrome P-450 dependent enzymes are involved in (S)-stylopine biosynthesis. *Phytochemistry, 30*(9), 2953-2961. doi: http://dx.doi.org/10.1016/S0031-9422(00)98230-X

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. [research-article]. *Journal of the Royal Statistical Society. Series B (Methodological)*(1), 289. doi: 10.2307/2346101

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., . . . Hannah, M. A. (2011). High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. *Plant J, 67*(5), 869-884. doi: 10.1111/j.1365-313X.2011.04640.x

Choi, K. B., Morishige, T., Shitan, N., Yazaki, K., & Sato, F. (2002). Molecular cloning and characterization of coclaurine N-methyltransferase from cultured cells of Coptis japonica. *J Biol Chem, 277*(1), 830-835. doi: 10.1074/jbc.M106405200

De-Eknamkul, W., & Zenk, M. H. (1992). Purification and properties of 1,2-dehydroreticuline reductase from Papaver somniferum seedlings. *Phytochemistry, 31*(3), 813-821. doi: http://dx.doi.org/10.1016/0031-9422(92)80020-F

Desgagné-Penix, I., Farrow, S., Cram, D., Nowak, J., & Facchini, P. (2012). Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Molecular Biology, 79*(3), 295-313. doi: 10.1007/s11103-012-9913-2

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLoS ONE, 6*(10), e26683. doi: 10.1371/journal.pone.0026683

Dittrich, H., & Kutchan, T. M. (1991). Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. *Proc Natl Acad Sci U S A, 88*(22), 9969-9973.

Facchini, P. J., & De Luca, V. (1994). Differential and tissue-specific expression of a gene family for tyrosine/dopa decarboxylase in opium poppy. *J Biol Chem, 269*(43), 26684-26690.

Fisinger, U., Grobe, N., & Zenk, M. H. (2007). Thebaine synthase: a new enzyme in the morphine pathway in Papaver somniferum. *Natural Product Communications, 2*(3), 249-253.

Gesell, A., Rolf, M., Ziegler, J., Diaz Chavez, M. L., Huang, F. C., & Kutchan, T. M. (2009). CYP719B1 is salutaridine synthase, the C-C phenol-coupling enzyme of morphine biosynthesis in opium poppy. *J Biol Chem, 284*(36), 24432-24442. doi: 10.1074/jbc.M109.033373

Gongora-Castillo, E., & Buell, C. R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Natural Product Reports, 30*(4), 490-500.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. [10.1038/nbt.1883]. *Nat Biotech, 29*(7), 644-652. doi: http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html - supplementary-information

Grothe, T., Lenz, R., & Kutchan, T. M. (2001). Molecular characterization of the salutaridinol 7-O-acetyltransferase involved in morphine biosynthesis in opium poppy Papaver somniferum. *J Biol Chem, 276*(33), 30717-30723. doi: 10.1074/jbc.M102688200

Hagel, J. M., Beaudoin, G. A., Fossati, E., Ekins, A., Martin, V. J., & Facchini, P. J. (2012). Characterization of a flavoprotein oxidase from opium poppy catalyzing the final steps in sanguinarine and papaverine biosynthesis. *J Biol Chem, 287*(51), 42972-42983. doi: 10.1074/jbc.M112.420414

Hagel, J. M., & Facchini, P. J. (2010). Dioxygenases catalyze the O-demethylation steps of morphine biosynthesis in opium poppy. *Nat Chem Biol, 6*(4), 273-275. doi: 10.1038/nchembio.317

Higashi, Y., & Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. *Plant, Cell & Environment*, n/a-n/a. doi: 10.1111/pce.12069

Hirata, K., Poeaknapo, C., Schmidt, J., & Zenk, M. H. (2004). 1,2-Dehydroreticuline synthase, the branch point enzyme opening the morphinan biosynthetic pathway. *Phytochemistry, 65*(8), 1039-1046. doi: 10.1016/j.phytochem.2004.02.015

Iseli, C., Jongeneel, C. V., & Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, 138-148.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*, 559-559. doi: 10.1186/1471-2105-9-559

Lee, E. J., & Facchini, P. (2010). Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. *Plant Cell, 22*(10), 3489-3503. doi: 10.1105/tpc.110.077958

Lee, E. J., & Facchini, P. J. (2011). Tyrosine aminotransferase contributes to benzylisoquinoline alkaloid biosynthesis in opium poppy. *Plant Physiol, 157*(3), 1067-1078. doi: 10.1104/pp.111.185512

Lenz, R., & Zenk, M. H. (1994). Closure of the oxide bridge in morphine biosynthesis. *Tetrahedron Letters, 35*(23), 3897-3900. doi: http://dx.doi.org/10.1016/S0040-4039(00)76696-2

Li, A., & Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics (Oxford, England), 23*(2), 222-231.

Li, B., & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*(1), 323.

Liscombe, D. K., & Facchini, P. J. (2007). Molecular cloning and characterization of tetrahydroprotoberberine cis-N-methyltransferase, an enzyme involved in alkaloid biosynthesis in opium poppy. *J Biol Chem, 282*(20), 14741-14751. doi: 10.1074/jbc.M611908200

Mao, L., Begum, D., Chuang, H. W., Budiman, M. A., Szymkowiak, E. J., Irish, E. E., & Wing, R. A. (2000). JOINTLESS is a MADS-box gene controlling tomato flower abscission zone development. *Nature, 406*(6798), 910-913. doi: 10.1038/35022611

Mishra, S., Triptahi, V., Singh, S., Phukan, U. J., Gupta, M. M., Shanker, K., & Shukla, R. K. (2013). Wound Induced Tanscriptional Regulation of Benzylisoquinoline Pathway and Characterization of Wound Inducible PsWRKY Transcription Factor from <italic>Papaver somniferum</italic>. *PLoS ONE, 8*(1), e52784. doi: 10.1371/journal.pone.0052784

Morishige, T., Tsujita, T., Yamada, Y., & Sato, F. (2000). Molecular characterization of the S-adenosyl-L-methionine:3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase involved in isoquinoline alkaloid biosynthesis in Coptis japonica. *J Biol Chem, 275*(30), 23398-23405. doi: 10.1074/jbc.M002439200

Movahedi, S., Van Bel, M., Heyndrickx, K. S., & Vandepoele, K. (2012). Comparative co-expression analysis in plant biology. *Plant Cell Environ, 35*(10), 1787-1798. doi: 10.1111/j.1365-3040.2012.02517.x

Pauli, H. H., & Kutchan, T. M. (1998). Molecular cloning and functional heterologous expression of two alleles encoding (S)-N-methylcoclaurine 3'-hydroxylase (CYP80B1), a new methyl jasmonate-inducible cytochrome P-450-dependent mono-oxygenase of benzylisoquinoline alkaloid biosynthesis. *Plant J, 13*(6), 793-801.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi: 10.1093/bioinformatics/btp616

Saito, K., Hirai, M. Y., & Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics – 'majority report by precogs'. *Trends in Plant Science, 13*(1), 36-43. doi: http://dx.doi.org/10.1016/j.tplants.2007.10.006

Schilmiller, A. L., Pichersky, E., & Last, R. L. (2012). Taming the hydra of specialized metabolism: how systems biology and comparative approaches are revolutionizing plant biochemistry. *Current Opinion in Plant Biology, 15*(3), 338-344. doi: http://dx.doi.org/10.1016/j.pbi.2011.12.005

Takemura, T., Ikezawa, N., Iwasa, K., & Sato, F. (2012). Molecular cloning and characterization of a cytochrome P450 in sanguinarine biosynthesis from Eschscholzia californica cells. *Phytochemistry*. doi: 10.1016/j.phytochem.2012.02.013

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., . . . Saito, K. (2008). Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis. *Plant Cell, 20*(8), 2160-2176. doi: 10.1105/tpc.108.058040

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics, 17*(9), 847-848. doi: 10.1093/bioinformatics/17.9.847

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications In Genetics And Molecular Biology, 4*, Article17-Article17.

Ziegler, J., Voigtlander, S., Schmidt, J., Kramell, R., Miersch, O., Ammer, C., . . . Kutchan, T. M. (2006). Comparative transcript and alkaloid profiling in Papaver species identifies a short chain dehydrogenase/reductase involved in morphine biosynthesis. *Plant J, 48*(2), 177-192. doi: 10.1111/j.1365-313X.2006.02860.x

APPENDIX A

GOSLIM CLASSIFICATION CHARTS

The following pages contain alphabetically arranged GOSlim profiles for each of the thirty-four identified transcriptional modules. Module-function specialization can be clearly seen in most modules. In each case the GOSlim profiles were generated by mapping the Gene Ontology terms from the aforementioned InterProScan output onto the sequences contained in each network module. These Gene Ontology terms were then reduced to GOSlim terms though the goStats R package. Relative percentages of GOSlim classifications were then used to form the charts shown. Tables A.1 through A.3 contain false discovery rate adjusted q-values for a test of enrichment for each GOSlim term compared to the background whole-transcriptome GOSlim counts.

Table A.1: GOSlim Biological Process

Enrichment statistics for GOSlim terms in the biological process category. P-values were generated using a Fisher's exact test for the term frequency within a module vs. the term frequency in all transcripts within the network. Shown are false discovery rate corrected q-values.

Term	black	blue	brown	cyan	darkgreen	darkgrey	darkmagenta	darkolivegreen
carbohydrate metabolic process	2.39E-03	7.07E-02	2.17E-05	4.39E-03	5.23E-01	6.76E-07	9.22E-05	2.02E-02
nucleobase-containing compound metabolic process	8.26E-02	6.85E-07	2.54E-03	2.99E-01	4.60E-01	2.03E-07	2.42E-11	7.83E-02
translation	4.06E-05	2.64E-02	8.83E-31	2.61E-07	4.16E-04	5.88E-04	5.86E-02	5.23E-01
protein modification process	2.14E-21	8.15E-14	1.29E-36	9.31E-04	1.30E-01	5.50E-05	6.11E-02	4.59E-01
lipid metabolic process	4.11E-05	5.09E-01	4.33E-01	6.01E-02	3.93E-01	4.84E-01	2.37E-01	3.65E-01
transport	8.95E-02	7.24E-02	1.47E-08	3.30E-05	4.53E-01	1.26E-02	4.12E-02	4.62E-01
cell cycle	3.11E-01	4.60E-01	5.23E-01	1.96E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01
cell communication	4.85E-01	6.82E-02	1.39E-07	5.52E-03	9.28E-02	4.76E-02	2.02E-03	5.23E-01
multicellular organismal development	5.23E-01	5.97E-02	8.33E-02	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01
photosynthesis	4.32E-07	6.54E-14	1.50E-14	6.45E-04	2.17E-02	2.13E-02	4.14E-01	3.11E-01
cellular component organization	1.89E-04	1.00E-09	5.23E-08	1.67E-04	1.06E-01	9.75E-07	5.58E-02	3.43E-03
secondary metabolic process	5.23E-01	1.79E-01	3.02E-01	2.56E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01

Table A.1 continued

Term	darkorange	darkred	darkturquoise	green	greenyellow	grey60	lightcyan	lightgreen
carbohydrate metabolic process	2.42E-01	3.70E-05	4.46E-06	2.67E-01	2.03E-01	9.57E-04	7.31E-02	7.47E-03
nucleobase-containing compound metabolic process	2.60E-10	6.76E-02	8.46E-14	1.68E-07	1.70E-06	4.60E-02	2.52E-05	3.43E-03
translation	2.09E-05	9.82E-06	1.44E-05	2.68E-91	1.83E-03	4.01E-03	2.46E-03	4.56E-01
protein modification process	6.11E-10	5.23E-01	1.33E-03	3.40E-01	3.31E-10	3.34E-01	5.23E-01	7.89E-02
lipid metabolic process	2.75E-03	4.38E-01	1.62E-02	1.11E-02	1.98E-03	2.91E-01	2.05E-02	1.22E-01
transport	5.23E-01	3.39E-01	5.18E-10	7.11E-11	1.07E-17	1.79E-01	3.98E-01	2.43E-21
cell cycle	5.23E-01	5.23E-01	9.88E-14	2.36E-01	3.09E-01	5.23E-01	3.59E-01	5.23E-01
cell communication	5.23E-01	4.53E-03	6.27E-04	2.76E-02	1.42E-13	1.52E-03	2.02E-03	2.44E-02
multicellular organismal development	1.44E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	3.87E-02	1.74E-02
photosynthesis	3.32E-01	3.15E-02	3.05E-02	1.63E-02	1.08E-137	2.68E-03	2.58E-03	2.98E-02
cellular component organization	1.68E-02	5.31E-02	4.15E-01	4.24E-01	1.87E-03	4.64E-02	2.51E-03	3.68E-01
secondary metabolic process	1.58E-01	5.23E-01	5.23E-01	3.96E-01	5.23E-01	5.23E-01	5.23E-01	1.56E-01

Table A.1 continued

Term	lightyellow	magenta	midnightblue	orange	Pale turquoise	pink	purple	red
carbohydrate metabolic process	2.66E-05	5.23E-01	2.04E-02	1.55E-12	5.00E-09	4.84E-02	2.68E-01	2.62E-02
nucleobase-containing compound metabolic process	6.05E-03	9.73E-02	2.74E-05	1.35E-46	3.87E-01	4.05E-01	1.28E-18	4.09E-01
translation	4.90E-01	1.31E-14	3.09E-03	3.59E-10	3.09E-01	2.55E-03	3.89E-17	2.27E-17
protein modification process	3.49E-03	4.74E-01	8.19E-02	7.08E-16	4.09E-01	1.04E-01	1.04E-02	1.54E-01
lipid metabolic process	1.48E-01	5.67E-04	2.05E-02	2.60E-03	2.32E-01	1.35E-01	4.29E-01	5.36E-02
transport	5.01E-02	8.34E-10	1.94E-14	1.02E-35	1.87E-04	4.95E-01	1.75E-11	1.29E-06
cell cycle	5.23E-01	4.28E-01	5.23E-01	5.23E-01	5.23E-01	3.07E-01	3.08E-01	3.48E-01
cell communication	1.92E-03	3.12E-01	4.33E-01	2.33E-03	1.66E-01	2.04E-04	2.06E-02	3.66E-01
multicellular organismal development	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	3.31E-01	5.23E-01
photosynthesis	1.75E-01	9.56E-06	1.35E-03	5.54E-06	4.12E-01	2.00E-02	2.47E-01	1.70E-01
cellular component organization	1.84E-17	1.09E-01	4.99E-01	6.38E-73	7.98E-02	1.10E-06	2.40E-01	2.30E-08
secondary metabolic process	1.46E-01	4.06E-04	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01

Table A.1 continued

Term	royalblue	saddlebrown	salmon	sienna3	skyblue	steelblue
carbohydrate metabolic process	3.01E-10	1.46E-01	9.26E-14	2.21E-01	4.98E-01	3.87E-02
nucleobase-containing compound metabolic process	5.34E-02	1.27E-01	4.53E-01	2.78E-05	2.94E-01	1.37E-03
translation	4.27E-01	2.88E-02	5.47E-07	2.31E-01	4.17E-02	1.01E-02
protein modification process	6.10E-04	1.88E-05	4.68E-09	4.26E-01	5.42E-04	7.27E-02
lipid metabolic process	2.46E-02	1.73E-01	4.32E-01	4.12E-01	1.29E-01	8.86E-03
transport	4.00E-04	8.60E-07	1.25E-02	7.72E-07	5.50E-04	9.91E-03
cell cycle	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01
cell communication	3.77E-02	1.27E-01	1.10E-01	3.08E-01	4.16E-01	8.81E-05
multicellular organismal development	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01
photosynthesis	1.52E-02	4.24E-01	1.29E-03	5.23E-01	1.68E-10	2.29E-01
cellular component organization	1.76E-01	4.13E-02	5.56E-05	2.25E-01	7.28E-02	1.43E-02
secondary metabolic process	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01	5.23E-01

Table A.1 continued

Term	tan	turquoise	violet	white	yellow
carbohydrate metabolic process	3.42E-07	4.40E-01	6.81E-04	8.50E-02	4.01E-01
nucleobase-containing compound metabolic process	1.92E-01	2.61E-01	5.23E-02	1.07E-06	5.11E-01
translation	2.85E-03	2.19E-56	7.70E-02	8.92E-08	4.39E-13
protein modification process	1.52E-07	9.03E-27	5.64E-02	1.94E-08	1.09E-08
lipid metabolic process	1.36E-02	2.69E-02	7.72E-08	2.54E-01	5.00E-07
transport	4.56E-09	1.62E-01	1.59E-02	6.99E-02	3.17E-03
cell cycle	4.12E-01	8.12E-02	5.23E-01	5.23E-01	1.83E-01
cell communication	5.81E-03	3.41E-04	4.52E-07	4.39E-02	5.05E-02
multicellular organismal development	5.23E-01	2.16E-01	5.23E-01	5.23E-01	5.23E-01
photosynthesis	1.15E-05	9.37E-03	4.09E-01	3.96E-01	6.27E-11
cellular component organization	3.16E-01	7.58E-15	3.66E-01	3.13E-01	2.53E-07
secondary metabolic process	5.23E-01	1.45E-01	5.23E-01	5.23E-01	1.17E-01

Table A.2: GOSlim Molecular Function

Enrichment statistics for GOSlim terms in the molecular function category. P-values were generated using a Fisher's exact test for the term frequency within a module vs. the term frequency in all transcripts within the network. Shown are false discovery rate corrected q-values.

Term	black	blue	brown	cyan	darkgreen	darkgrey	darkmagenta	darkolivegreen	darkorange
nucleotide binding	4.00E-01	8.94E-09	3.96E-01	4.28E-03	1.32E-02	1.14E-02	5.53E-03	5.10E-02	4.10E-01
protein binding	4.46E-18	2.35E-17	2.15E-10	4.77E-01	3.31E-01	1.68E-01	1.41E-01	7.56E-08	2.16E-01
nucleic acid binding	4.47E-09	1.28E-44	1.32E-02	6.13E-02	1.20E-01	5.57E-08	2.89E-01	2.30E-03	2.01E-07
motor activity	5.50E-03	4.54E-02	1.25E-04	1.72E-05	2.73E-02	3.84E-02	3.34E-05	2.73E-01	3.89E-02
catalytic activity	9.28E-20	8.35E-74	4.32E-01	5.22E-02	3.63E-01	1.84E-01	7.06E-02	4.77E-01	2.91E-01
signal transducer activity	4.77E-01	1.10E-06	8.35E-11	2.42E-01	1.17E-11	1.97E-02	8.36E-18	1.87E-01	2.04E-02
carbohydrate binding	1.47E-03	1.68E-05	1.57E-01	3.44E-01	2.02E-01	3.43E-01	3.67E-01	3.68E-01	3.48E-01
structural molecule activity	1.30E-02	7.05E-02	1.12E-19	7.00E-02	1.35E-03	2.29E-02	1.51E-01	1.01E-01	1.82E-03
transporter activity	2.84E-01	3.64E-02	1.23E-05	2.28E-01	7.68E-02	4.32E-01	5.13E-02	2.59E-01	8.37E-02
lipid binding	1.59E-01	1.01E-05	2.30E-03	2.99E-44	1.88E-02	1.12E-02	3.69E-01	3.69E-01	5.38E-03
oxygen binding	4.77E-01	3.62E-01	3.61E-01	3.69E-14	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01
enzyme regulator activity	3.46E-06	1.01E-01	5.54E-02	4.49E-01	3.83E-03	2.56E-03	1.43E-01	7.23E-02	4.40E-24

Table A.2 continued

Term	darkred	darkturquoise	green	greenyellow	grey60	lightcyan	lightgreen	lightyellow	magenta
nucleotide binding	4.21E-01	9.74E-20	4.09E-01	4.80E-19	9.25E-05	1.80E-01	1.55E-01	1.87E-02	3.21E-01
protein binding	4.59E-02	1.71E-01	2.31E-02	1.80E-14	1.77E-03	1.32E-12	1.86E-18	8.59E-14	2.94E-01
nucleic acid binding	2.04E-03	2.14E-12	4.67E-15	1.38E-01	2.24E-25	3.60E-01	1.14E-01	4.29E-36	3.18E-01
motor activity	1.02E-02	8.15E-78	4.07E-01	9.88E-02	1.84E-03	2.26E-07	9.82E-02	4.04E-02	4.52E-01
catalytic activity	2.65E-02	1.12E-43	2.37E-01	2.22E-17	4.14E-07	4.67E-09	6.85E-11	3.08E-06	4.29E-01
signal transducer activity	2.82E-01	4.90E-04	1.01E-01	4.73E-08	1.31E-02	6.88E-02	2.70E-02	6.56E-02	4.30E-01
carbohydrate binding	1.96E-02	8.94E-02	1.88E-01	4.16E-01	1.72E-01	2.63E-01	2.78E-02	3.51E-01	1.83E-02
structural molecule activity	1.89E-02	4.00E-06	1.46E-99	3.27E-13	1.52E-03	1.75E-05	4.99E-02	2.16E-03	2.61E-02
transporter activity	3.98E-01	1.44E-13	4.77E-01	1.84E-05	4.62E-01	4.61E-01	1.19E-13	1.43E-01	2.09E-06
lipid binding	1.87E-01	3.70E-03	1.46E-03	4.28E-04	6.88E-02	2.71E-01	3.68E-01	3.51E-14	1.28E-06
oxygen binding	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01
enzyme regulator activity	4.77E-01	1.66E-03	5.14E-05	3.43E-11	2.98E-01	7.16E-05	4.77E-01	4.61E-02	2.09E-01

Table A.2 continued

Term	midnightblue	orange	paleturquoise	pink	purple	red	royalblue	saddlebrown	salmon
nucleotide binding	6.70E-02	8.19E-06	2.40E-02	1.38E-01	2.69E-01	9.34E-03	4.43E-01	4.77E-01	3.87E-18
protein binding	9.23E-02	2.42E-06	4.53E-17	4.48E-01	2.91E-01	1.28E-11	1.81E-01	1.19E-02	2.88E-02
nucleic acid binding	2.88E-01	4.86E-122	2.12E-01	4.86E-02	1.43E-17	5.82E-10	3.09E-01	1.15E-02	4.77E-03
motor activity	3.23E-03	9.29E-04	3.69E-01	2.18E-04	3.00E-01	4.07E-01	1.40E-02	1.45E-01	1.05E-04
catalytic activity	1.47E-01	6.23E-57	7.16E-04	9.32E-02	4.68E-01	2.62E-01	3.74E-01	2.16E-01	1.07E-06
signal transducer activity	2.52E-01	1.66E-03	2.73E-01	1.73E-03	1.79E-01	2.25E-06	8.99E-03	6.99E-02	4.48E-01
carbohydrate binding	8.73E-02	5.26E-02	2.05E-01	2.75E-01	4.08E-01	4.77E-01	4.24E-01	3.99E-01	4.02E-01
structural molecule activity	1.69E-01	1.50E-02	7.52E-02	4.11E-03	1.61E-08	1.95E-12	8.38E-02	2.80E-02	1.67E-08
transporter activity	8.90E-02	6.91E-07	1.71E-03	9.45E-02	1.13E-15	4.11E-03	2.70E-01	1.62E-06	4.07E-01
lipid binding	3.99E-01	1.97E-04	3.66E-01	3.72E-05	2.88E-01	3.74E-05	2.87E-02	2.05E-01	3.18E-04
oxygen binding	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01	4.77E-01
enzyme regulator activity	1.77E-01	2.68E-01	2.00E-01	1.02E-05	3.52E-17	1.72E-05	9.34E-04	3.83E-02	9.16E-15

Table A.2 continued

Term	sienna3	skyblue	steelblue	tan	turquoise	violet	white	yellow
nucleotide binding	8.27E-03	3.99E-01	4.63E-01	5.11E-02	1.31E-33	1.23E-01	4.15E-01	2.65E-06
protein binding	2.09E-04	3.88E-01	5.98E-02	3.72E-08	6.90E-03	1.66E-01	9.57E-02	9.30E-03
nucleic acid binding	3.78E-01	2.69E-02	1.53E-16	2.59E-15	6.88E-03	1.29E-02	2.51E-01	9.33E-08
motor activity	4.77E-01	7.42E-02	8.07E-07	4.10E-02	6.04E-05	3.65E-01	1.08E-01	8.84E-02
catalytic activity	6.57E-04	4.33E-02	9.15E-11	7.69E-06	7.10E-03	3.40E-01	1.71E-01	9.19E-03
signal transducer activity	4.77E-01	9.11E-02	1.53E-01	4.50E-01	1.20E-03	3.69E-01	5.53E-02	4.40E-01
carbohydrate binding	1.22E-01	4.05E-01	1.02E-01	1.92E-01	4.33E-01	3.47E-01	3.25E-01	9.18E-03
structural molecule activity	2.73E-01	1.50E-01	7.42E-03	8.47E-03	1.07E-02	1.99E-01	3.64E-11	2.07E-17
transporter activity	6.83E-05	1.95E-07	1.82E-01	8.72E-11	4.72E-01	3.28E-01	2.02E-01	6.31E-02
lipid binding	8.34E-03	4.08E-01	4.11E-01	8.38E-04	6.60E-16	3.55E-01	1.02E-01	1.73E-02
oxygen binding	4.77E-01	4.77E-01	4.77E-01	4.77E-01	2.64E-01	4.77E-01	4.77E-01	4.77E-01
enzyme regulator activity	3.70E-01	5.78E-03	1.40E-02	3.60E-02	1.39E-02	3.83E-01	1.27E-01	1.63E-03

Table A.3: GOSlim Cellular Component

Enrichment statistics for GOSlim terms in the cellular component category. P-values were generated using a Fisher's exact test for the term frequency within a module vs. the term frequency in all transcripts within the network. Shown are false discovery rate corrected q-values.

Term	black	blue	brown	cyan	darkgreen	darkgrey	darkmagenta	darkolivegreen
plasma membrane	7.12E-01	6.70E-01	5.14E-01	7.18E-01	7.97E-01	4.76E-01	7.97E-01	7.97E-01
mitochondrion	5.20E-01	1.87E-01	6.29E-01	7.97E-01	7.97E-01	8.41E-05	7.97E-01	7.97E-01
vacuole	7.97E-01	3.86E-01	7.97E-01	7.97E-01	7.97E-01	8.24E-02	7.97E-01	7.97E-01
extracellular region	5.11E-01	1.31E-03	7.97E-01	3.21E-06	1.74E-07	8.85E-02	7.97E-01	7.04E-01
cytosol	5.06E-01	7.58E-01	7.97E-01	5.10E-01	7.97E-01	7.15E-01	7.97E-01	7.97E-01
endoplasmic reticulum	1.01E-01	7.45E-01	1.92E-03	7.18E-01	7.97E-01	6.94E-01	6.24E-03	7.97E-01
peroxisome	9.51E-06	7.97E-01	6.71E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
ribosome	1.45E-02	2.10E-07	7.61E-11	2.49E-09	1.72E-01	6.54E-06	5.78E-01	4.55E-01
cell wall	6.34E-01	3.29E-01	7.70E-01	5.72E-01	7.97E-01	3.45E-01	7.97E-01	7.97E-01
nucleus	1.80E-02	1.87E-03	1.75E-16	1.22E-01	4.20E-01	3.86E-06	7.28E-01	3.15E-03
Golgi apparatus	7.16E-01	7.50E-01	7.97E-01	9.40E-05	7.97E-01	2.99E-07	7.97E-01	7.97E-01
cytoskeleton	7.48E-01	3.18E-05	9.30E-02	9.97E-02	6.04E-01	1.54E-02	7.97E-01	7.97E-01
plastid	5.09E-01	5.57E-01	6.95E-01	4.01E-01	7.97E-01	6.12E-01	7.97E-01	7.97E-01

Table A.3 continued

Term	darkorange	darkred	darkturquoise	green	greenyellow	grey60	lightcyan	lightgreen
plasma membrane	7.97E-01	7.97E-01	7.97E-01	6.96E-01	6.65E-01	5.62E-01	7.97E-01	7.97E-01
mitochondrion	7.97E-01	7.97E-01	7.97E-01	7.27E-01	1.09E-01	7.97E-01	7.97E-01	7.97E-01
vacuole	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
extracellular region	3.45E-04	7.03E-01	6.72E-02	4.24E-04	7.88E-05	4.88E-02	1.87E-01	6.54E-04
cytosol	7.97E-01	7.97E-01	7.97E-01	3.54E-02	2.26E-01	7.97E-01	7.97E-01	7.20E-01
endoplasmic reticulum	6.98E-01	7.97E-01	7.97E-01	5.10E-03	3.83E-01	6.11E-01	5.09E-01	2.72E-01
peroxisome	7.97E-01	7.97E-01	7.97E-01	7.06E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
ribosome	3.38E-02	5.84E-01	4.25E-03	5.28E-64	6.28E-02	1.47E-01	6.19E-05	1.70E-01
cell wall	8.69E-04	9.37E-02	7.11E-01	9.80E-03	5.82E-01	6.09E-01	8.09E-04	2.31E-01
nucleus	2.10E-04	6.34E-01	1.05E-10	1.05E-31	9.40E-10	5.47E-07	7.23E-01	3.94E-06
Golgi apparatus	7.97E-01	7.97E-01	7.97E-01	1.02E-01	4.01E-01	7.97E-01	7.97E-01	7.10E-01
cytoskeleton	5.17E-01	7.97E-01	2.13E-01	5.53E-01	7.41E-01	1.55E-01	2.23E-02	7.71E-01
plastid	7.97E-01	3.15E-01	7.97E-01	3.19E-01	1.51E-02	7.97E-01	7.97E-01	6.13E-01

Table A.3 continued

Term	lightyellow	magenta	midnightblue	orange	paleturquoise	pink	purple	red
plasma membrane	7.97E-01	6.12E-01	1.15E-01	5.03E-01	7.97E-01	7.97E-01	3.11E-01	7.97E-01
mitochondrion	6.15E-01	6.75E-01	5.77E-01	2.29E-01	7.97E-01	6.82E-01	2.55E-01	5.87E-01
vacuole	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
extracellular region	5.18E-03	5.86E-03	6.42E-01	3.38E-10	3.89E-01	5.81E-01	7.67E-01	4.40E-01
cytosol	1.71E-01	4.01E-01	7.10E-01	6.77E-01	7.97E-01	6.54E-01	7.97E-01	7.10E-01
endoplasmic reticulum	2.50E-01	7.24E-01	4.10E-01	3.85E-01	1.30E-01	6.15E-01	5.15E-01	4.16E-01
peroxisome	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
ribosome	3.50E-02	1.15E-13	1.91E-03	1.03E-11	6.97E-01	4.17E-01	3.08E-04	4.37E-03
cell wall	4.96E-01	7.64E-01	7.97E-01	1.31E-02	7.97E-01	7.22E-01	1.21E-03	7.42E-01
nucleus	3.14E-01	4.07E-01	7.57E-01	1.64E-40	2.85E-01	5.70E-01	6.39E-02	1.24E-03
Golgi apparatus	7.10E-01	7.33E-01	4.13E-01	7.38E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
cytoskeleton	1.65E-02	9.08E-11	4.19E-05	1.48E-02	4.94E-01	1.07E-01	1.02E-05	7.66E-01
plastid	6.12E-01	7.97E-01	7.11E-01	2.22E-01	7.97E-01	3.91E-01	4.92E-01	1.09E-01

Table A.3 continued

Term	royalblue	saddlebrown	salmon	sienna3	skyblue	steelblue
plasma membrane	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
mitochondrion	7.97E-01	7.97E-01	6.19E-01	7.97E-01	7.97E-01	7.97E-01
vacuole	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
extracellular region	5.06E-03	7.97E-01	3.76E-19	5.97E-01	6.19E-02	7.07E-01
cytosol	7.97E-01	7.97E-01	6.12E-01	7.97E-01	7.97E-01	7.97E-01
endoplasmic reticulum	7.97E-01	8.21E-02	6.17E-01	2.13E-14	7.97E-01	7.97E-01
peroxisome	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
ribosome	7.97E-01	5.78E-01	5.32E-09	3.48E-01	7.97E-01	4.63E-01
cell wall	7.97E-01	7.97E-01	2.45E-01	7.97E-01	2.17E-01	7.97E-01
nucleus	2.22E-01	4.59E-01	2.40E-01	3.95E-02	6.03E-01	3.95E-01
Golgi apparatus	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01	7.97E-01
cytoskeleton	5.08E-01	7.97E-01	7.95E-04	7.07E-01	7.10E-01	1.33E-06
plastid	7.97E-01	7.97E-01	6.13E-01	7.97E-01	7.97E-01	7.97E-01

Table A.3 continued

Term	tan	turquoise	violet	white	yellow
plasma membrane	7.06E-01	3.66E-02	7.97E-01	5.51E-01	7.42E-01
mitochondrion	1.21E-04	1.06E-01	7.97E-01	7.97E-01	3.45E-01
vacuole	1.05E-02	7.97E-01	7.97E-01	7.97E-01	7.97E-01
extracellular region	3.69E-03	1.10E-13	7.35E-01	4.63E-02	1.43E-01
cytosol	1.86E-01	9.60E-10	7.97E-01	7.97E-01	1.55E-01
endoplasmic reticulum	6.95E-01	3.96E-01	7.85E-02	7.17E-01	7.71E-01
peroxisome	7.97E-01	5.22E-01	7.97E-01	7.97E-01	7.97E-01
ribosome	5.42E-02	3.11E-24	1.72E-01	1.93E-11	6.96E-21
cell wall	7.97E-01	1.08E-01	6.28E-01	7.14E-01	4.53E-01
nucleus	5.97E-01	4.64E-18	1.18E-01	6.21E-07	3.57E-22
Golgi apparatus	5.31E-01	6.93E-01	7.97E-01	7.97E-01	7.97E-01
cytoskeleton	4.53E-01	5.74E-06	6.04E-01	1.47E-01	1.95E-03
plastid	6.13E-01	8.85E-06	7.97E-01	3.07E-04	1.65E-01

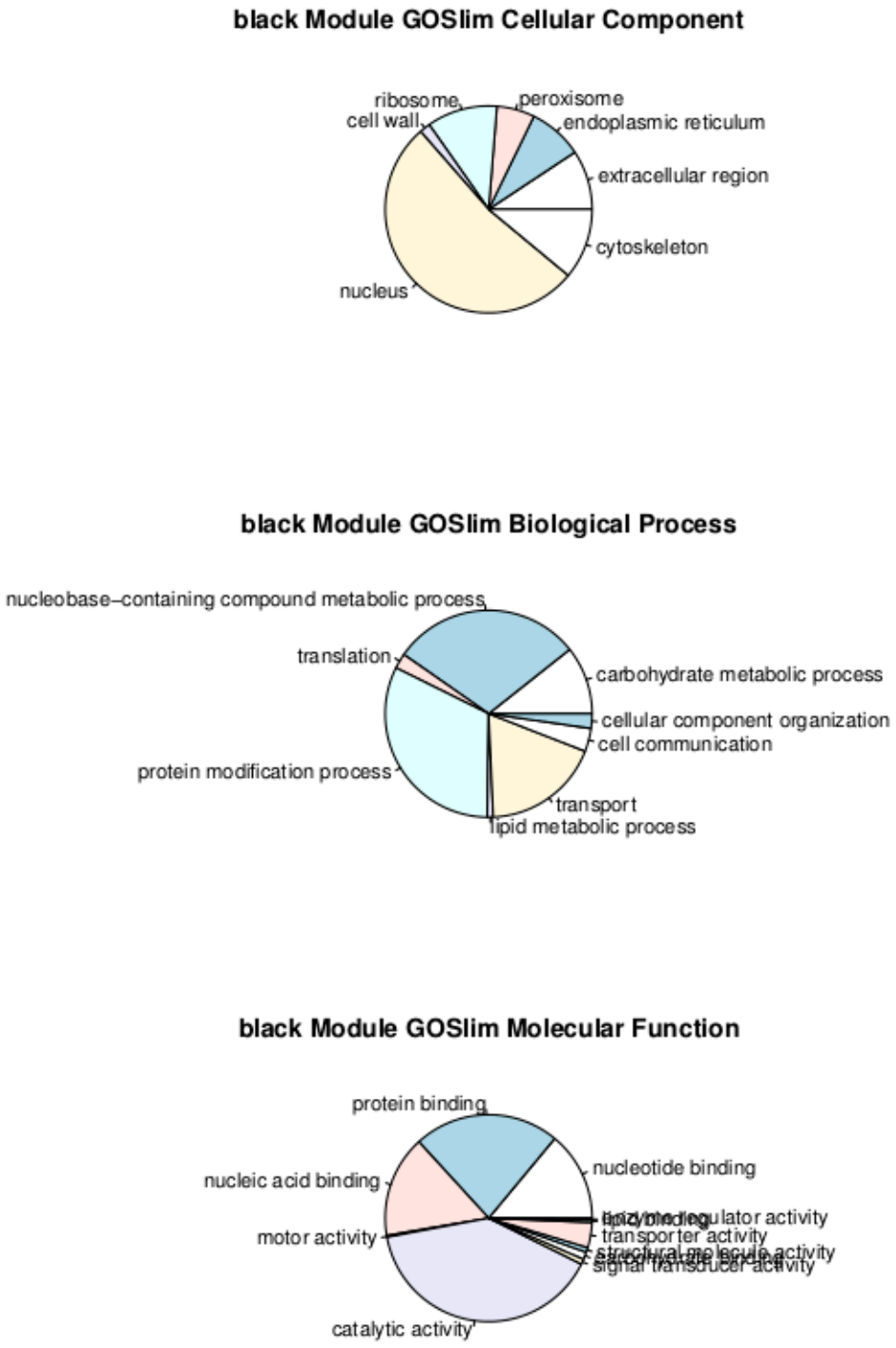
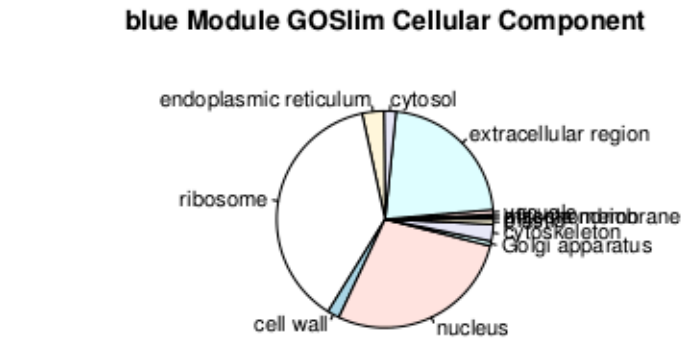
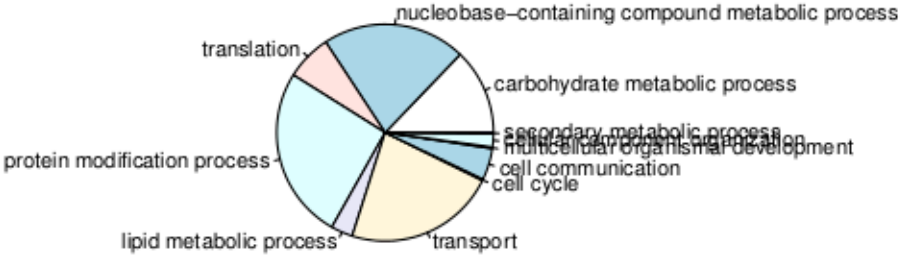


Figure A.1: Black module GOSlim classification distribution





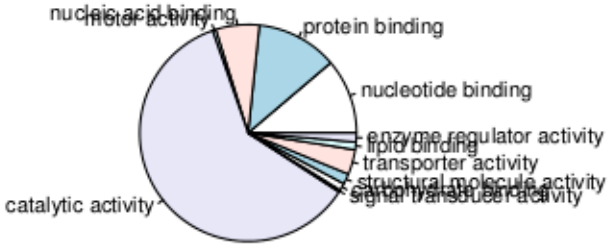


Figure A.2: Blue module GOSlim classification distribution

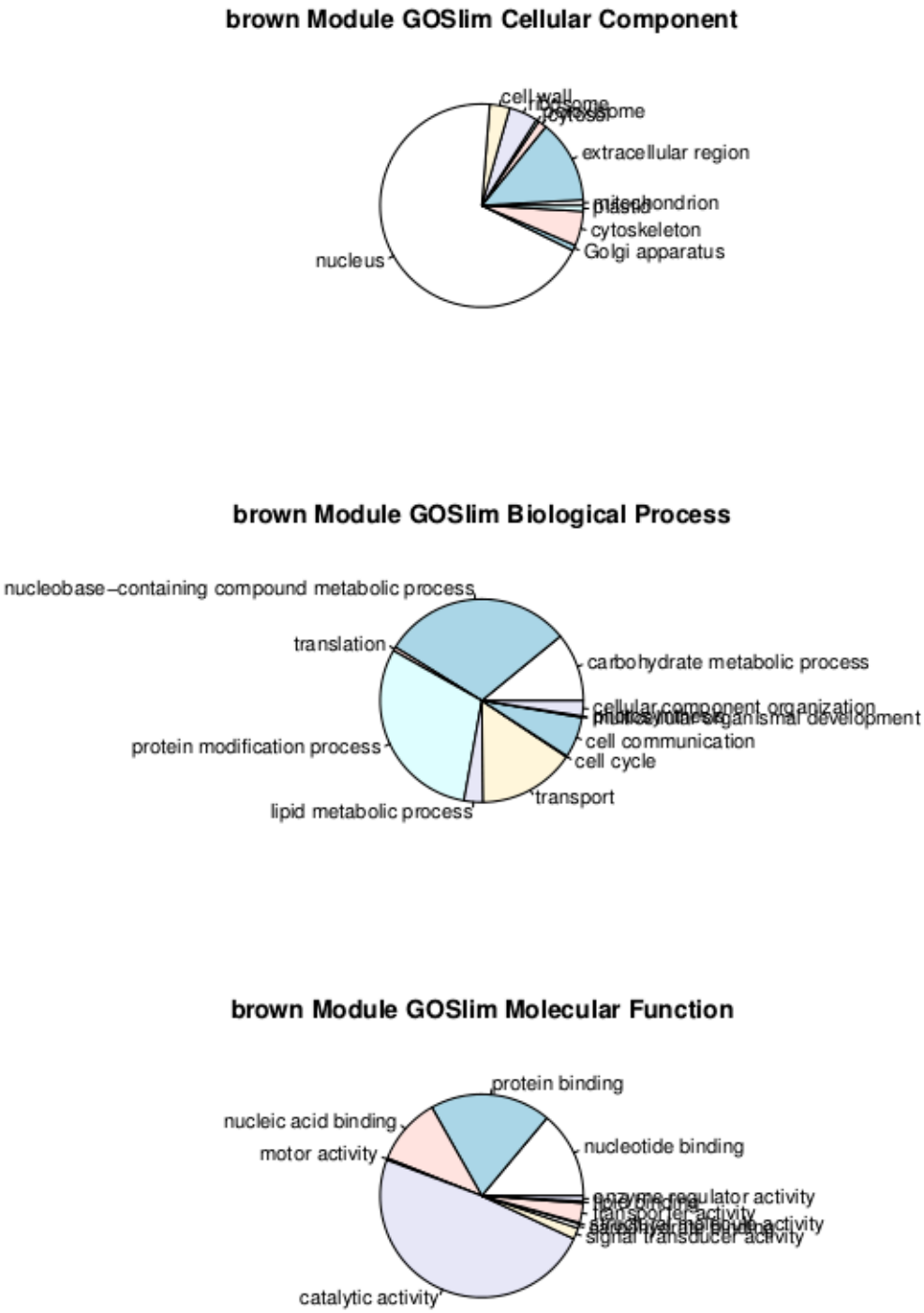


Figure A.3: Brown module GOSlim classification distribution

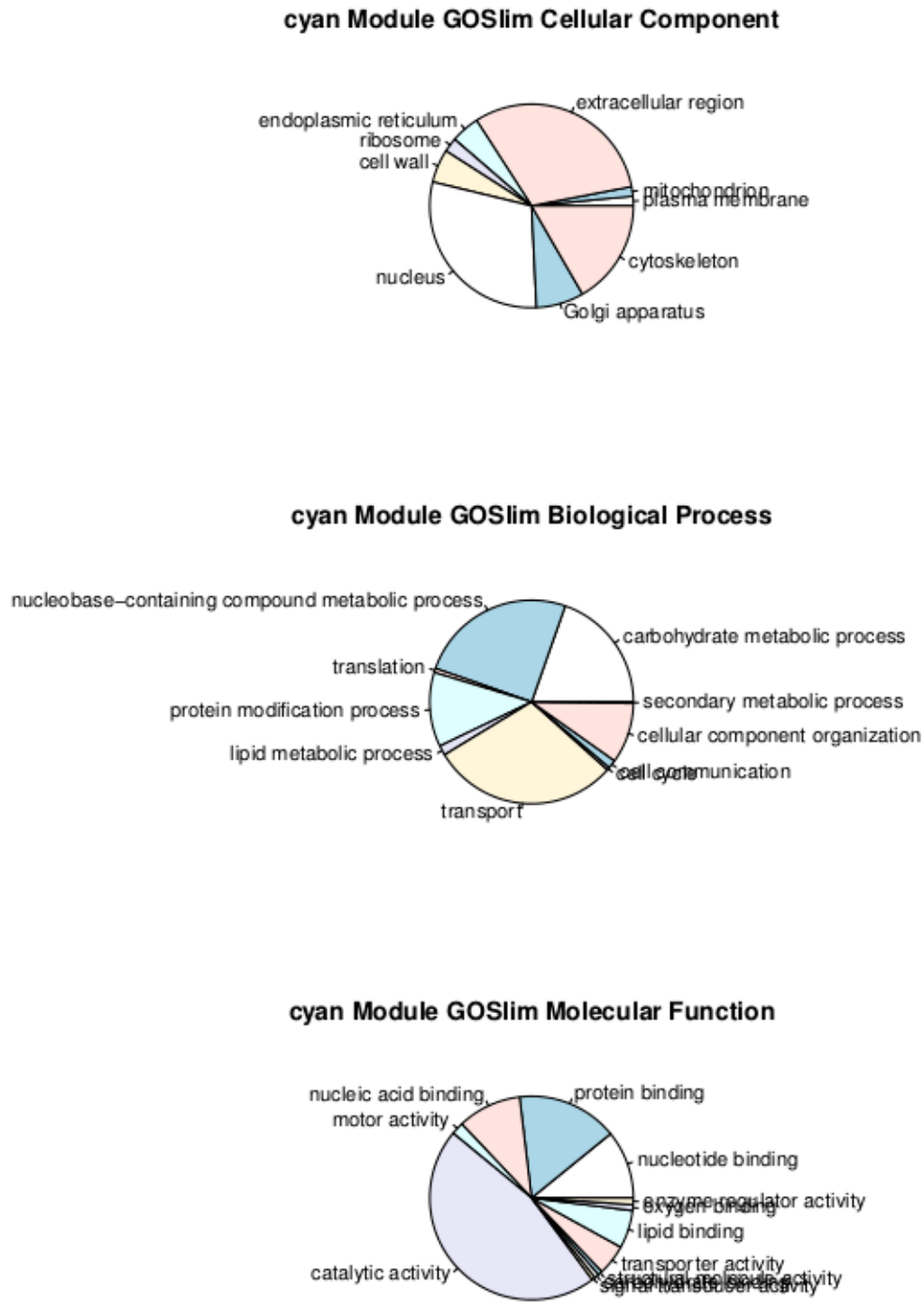


Figure A.4: Cyan module GOSlim classification distribution

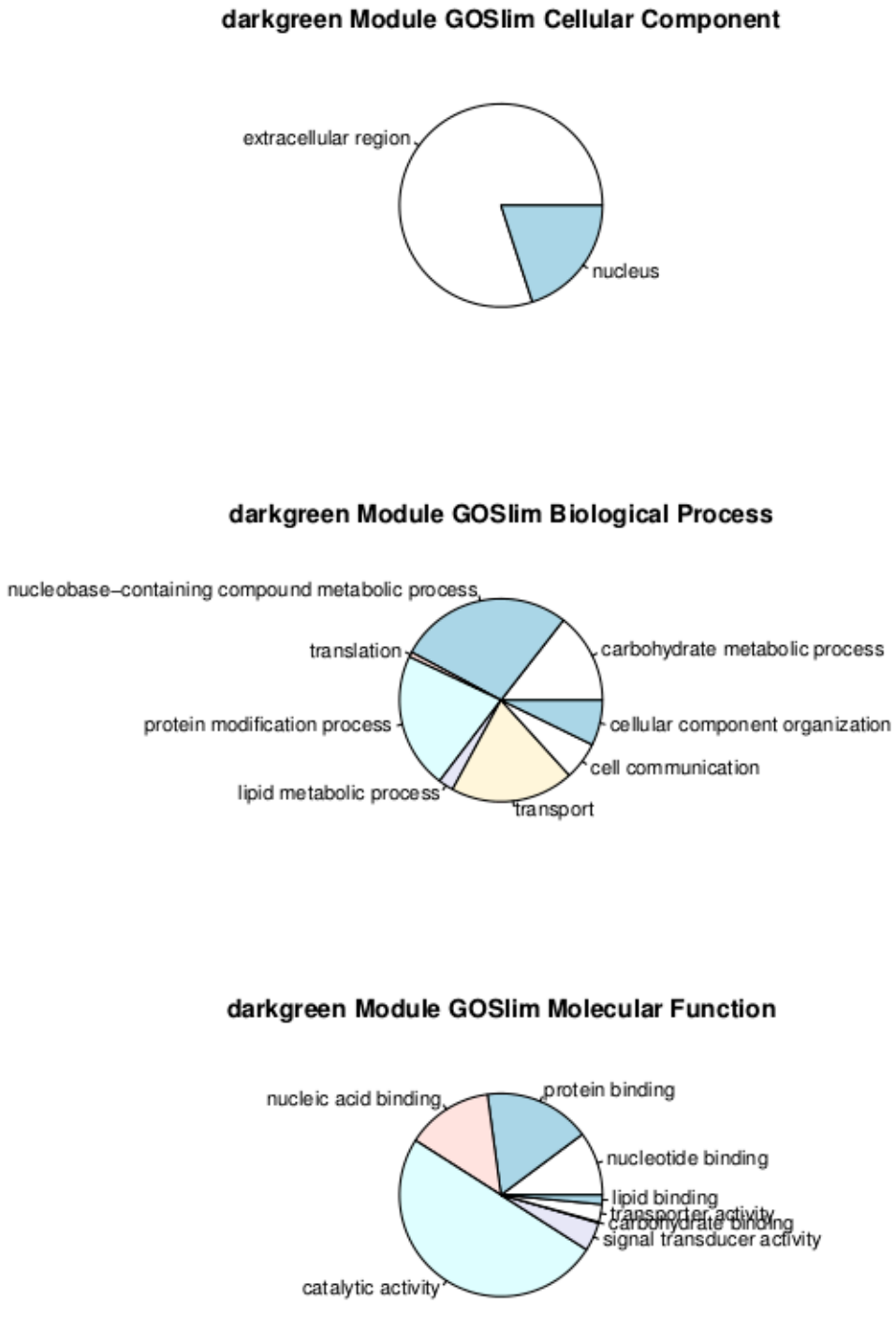


Figure A.5: Dark Green module GOSlim classification distribution

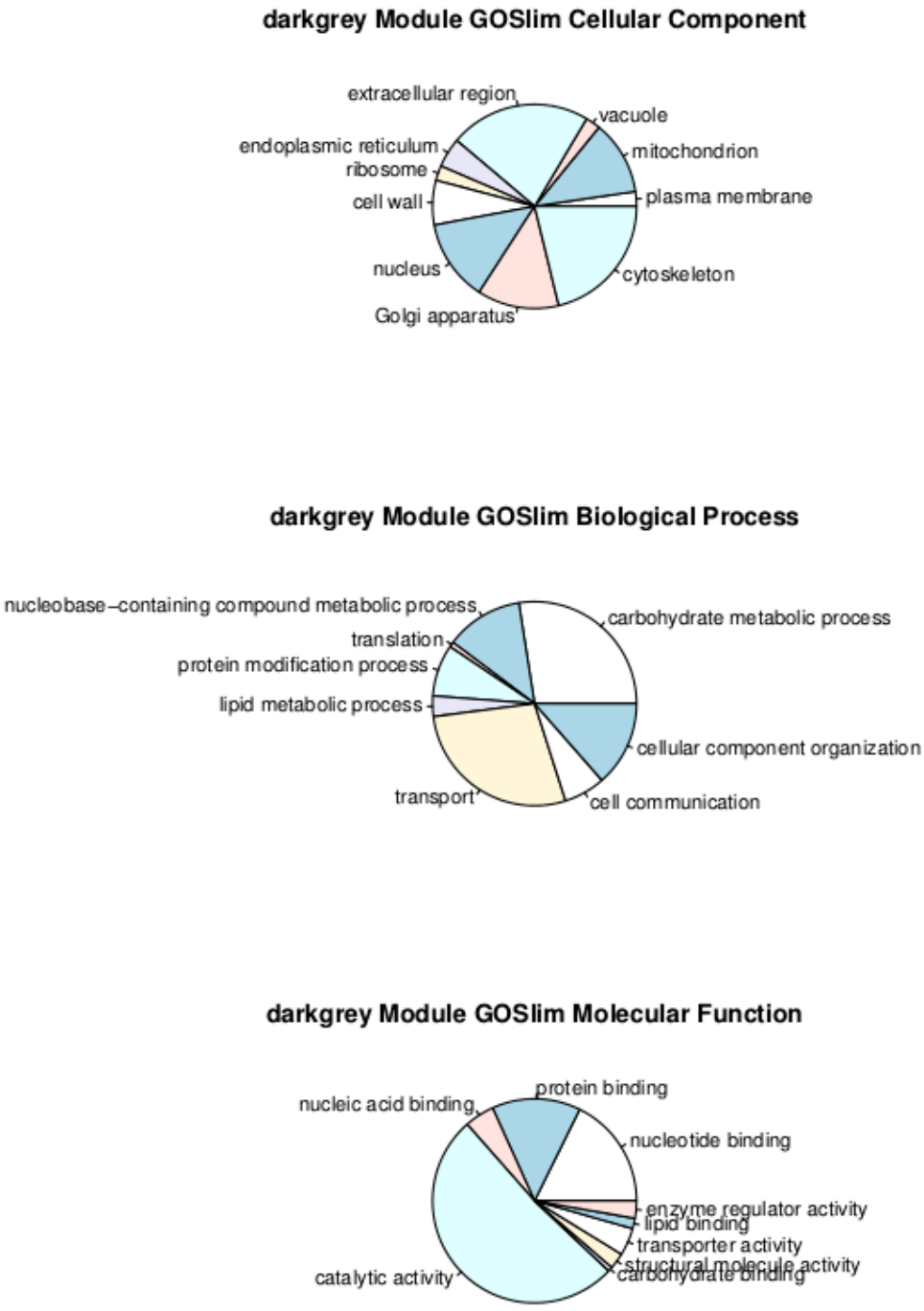


Figure A.6: Dark Grey module GOSlim classification distribution

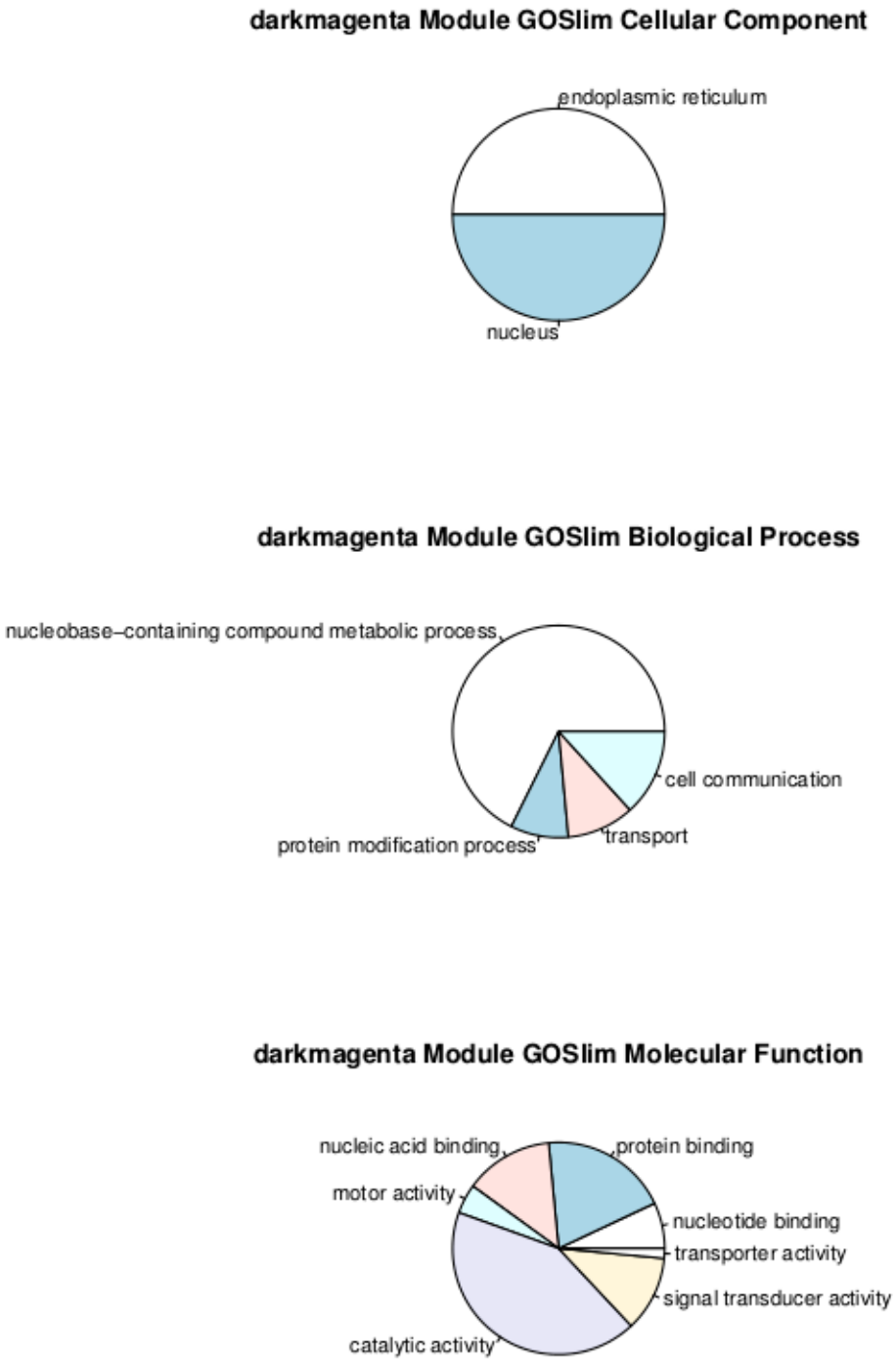


Figure A.7: Dark Magenta module GOSlim classification distribution

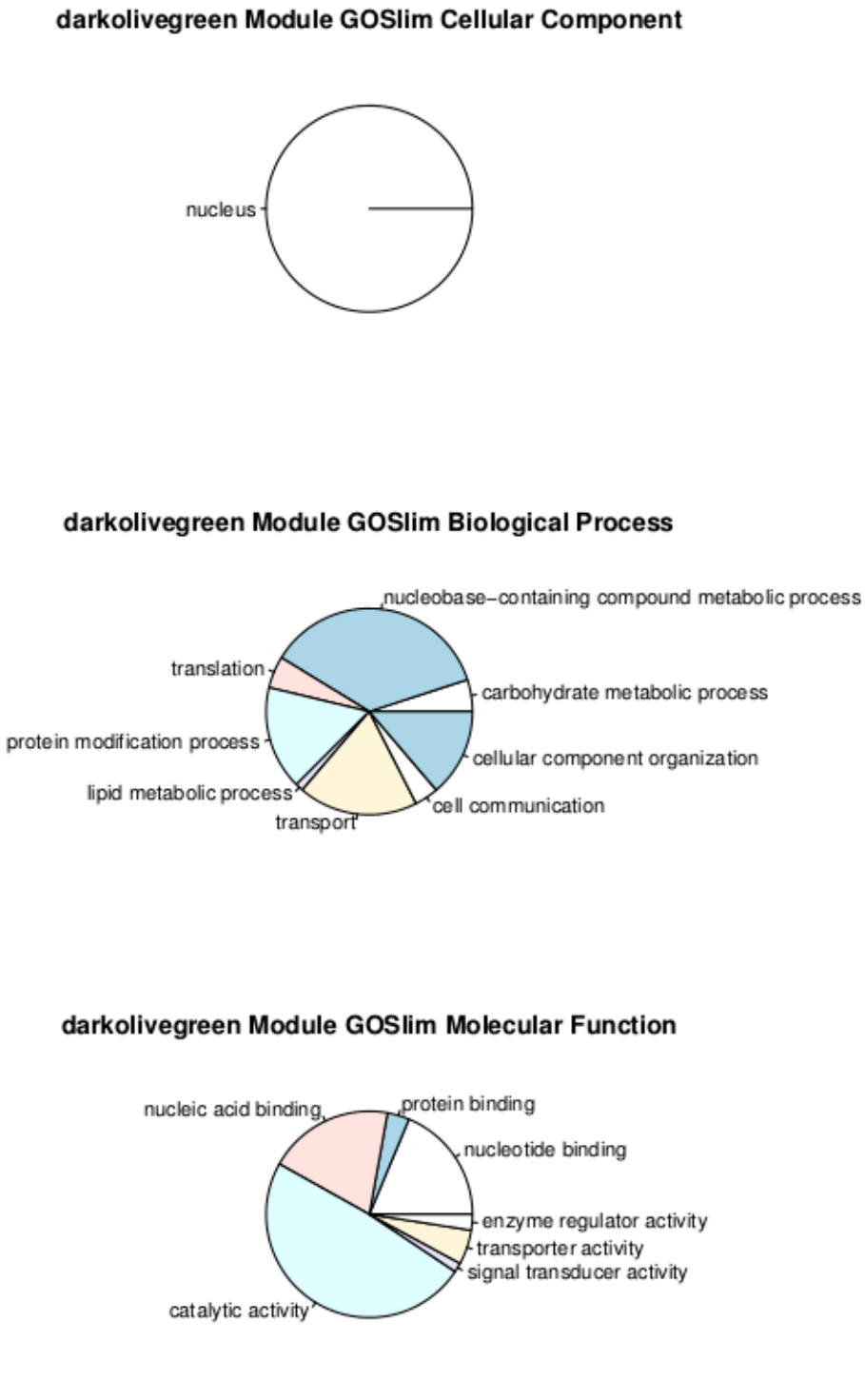
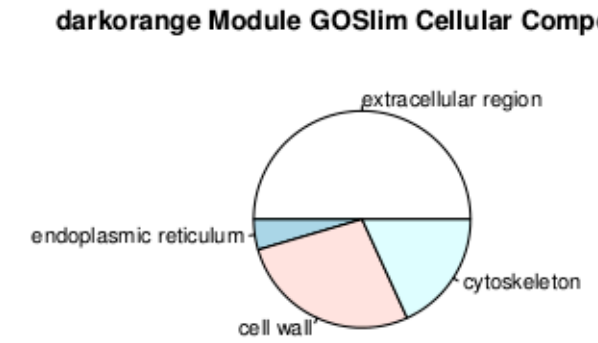
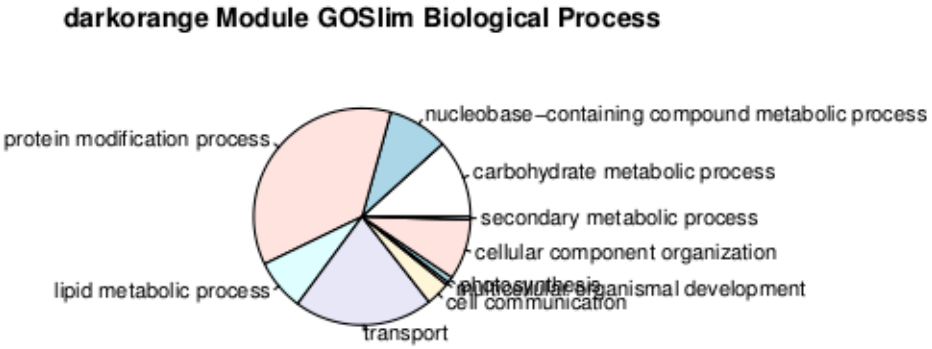


Figure A.8: Dark Olive Green module GOSlim classification distribution





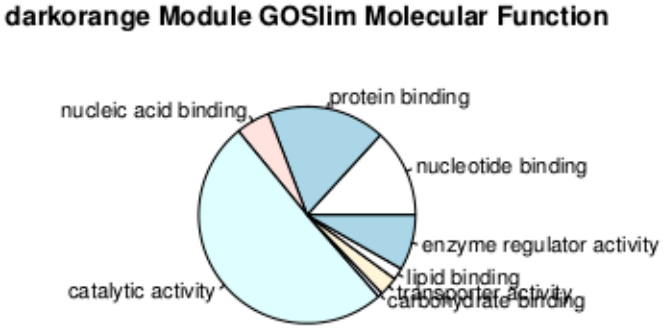


Figure A.9: Dark Orange module GOSlim classification distribution

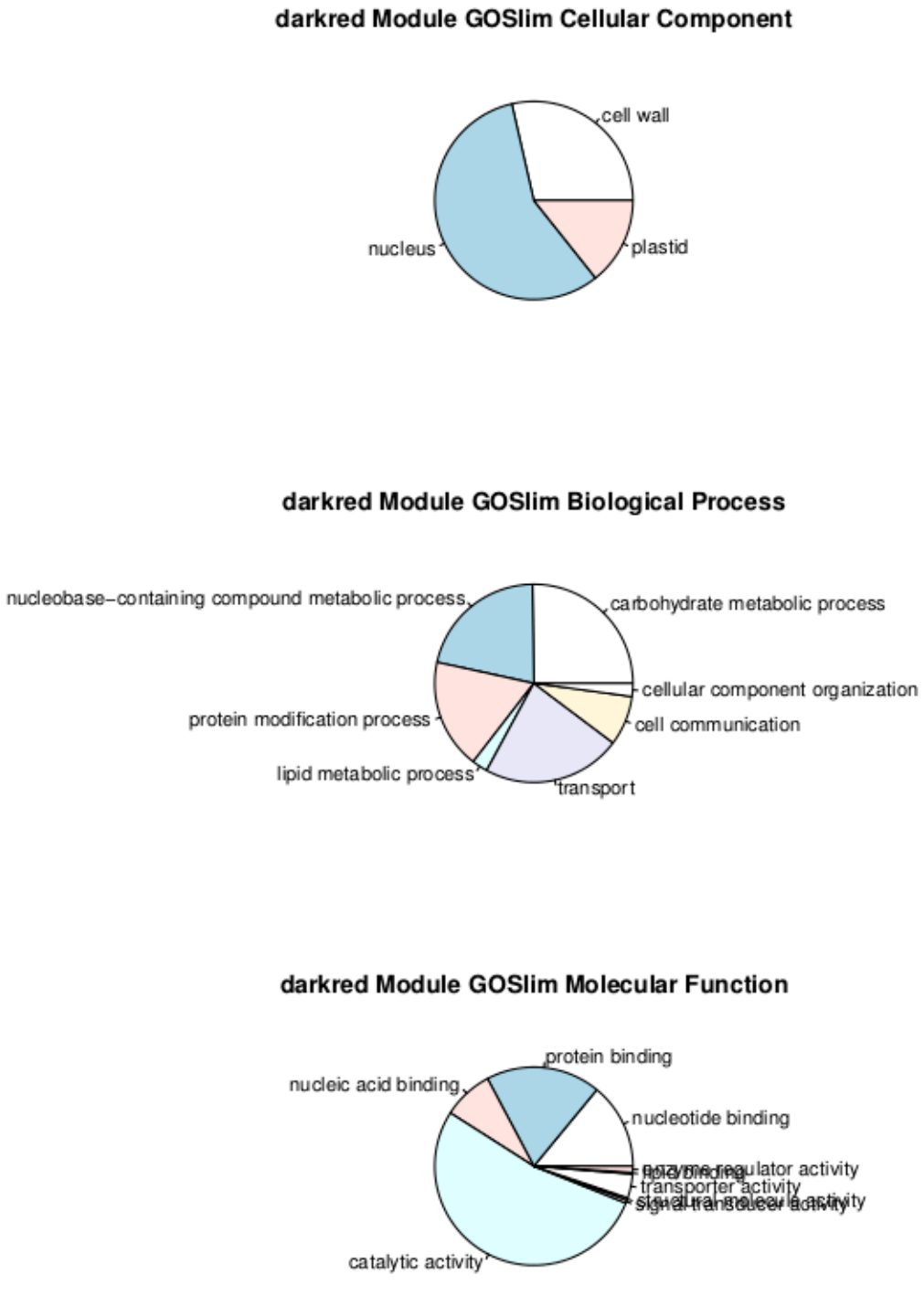


Figure A.10: Dark Red module GOSlim classification distribution

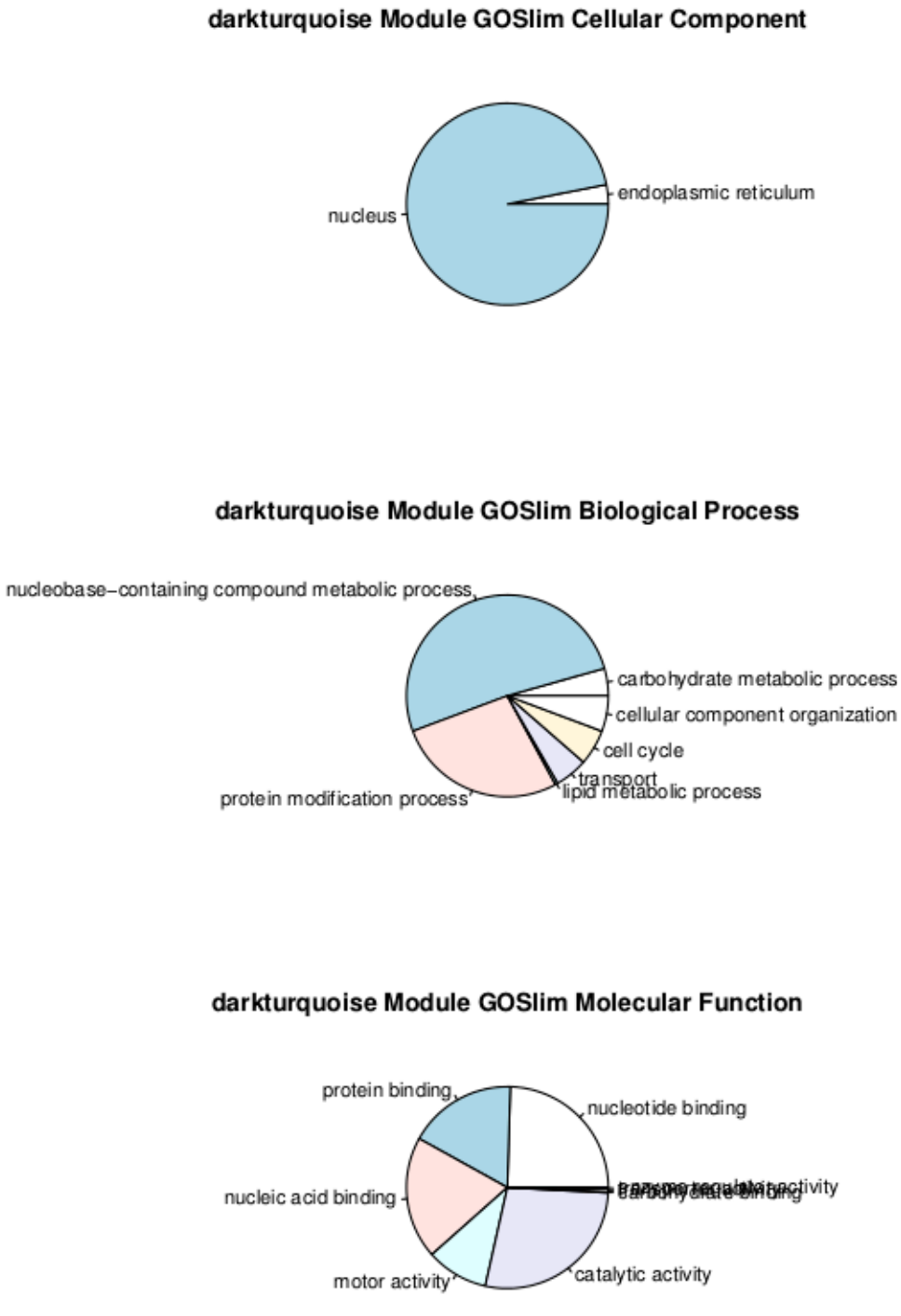
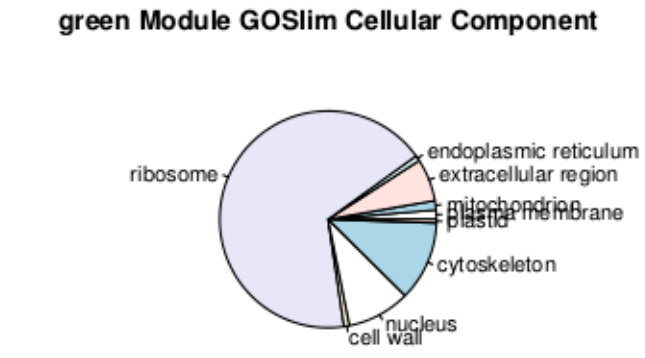
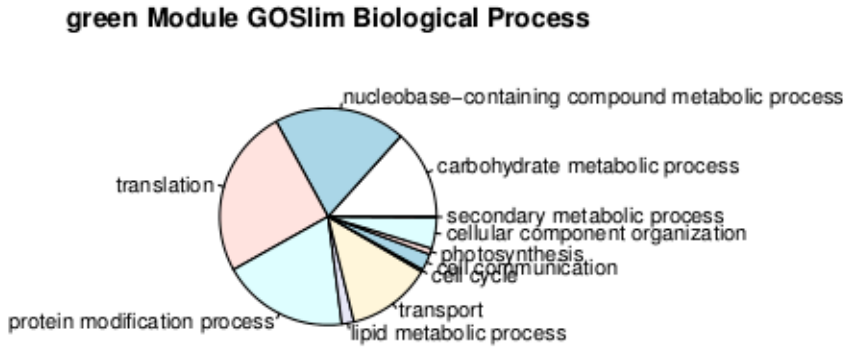


Figure A.11: Dark Turquoise module GOSlim classification distribution





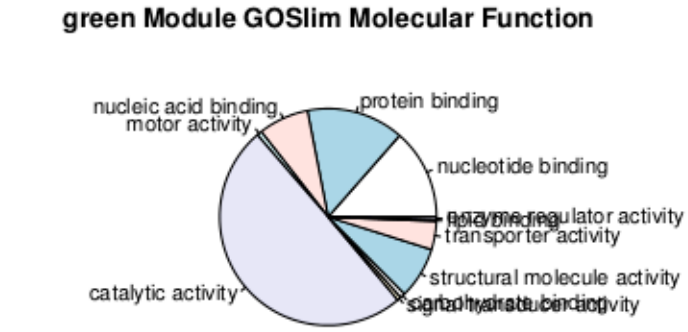


Figure A.12: Green module GOSlim classification distribution

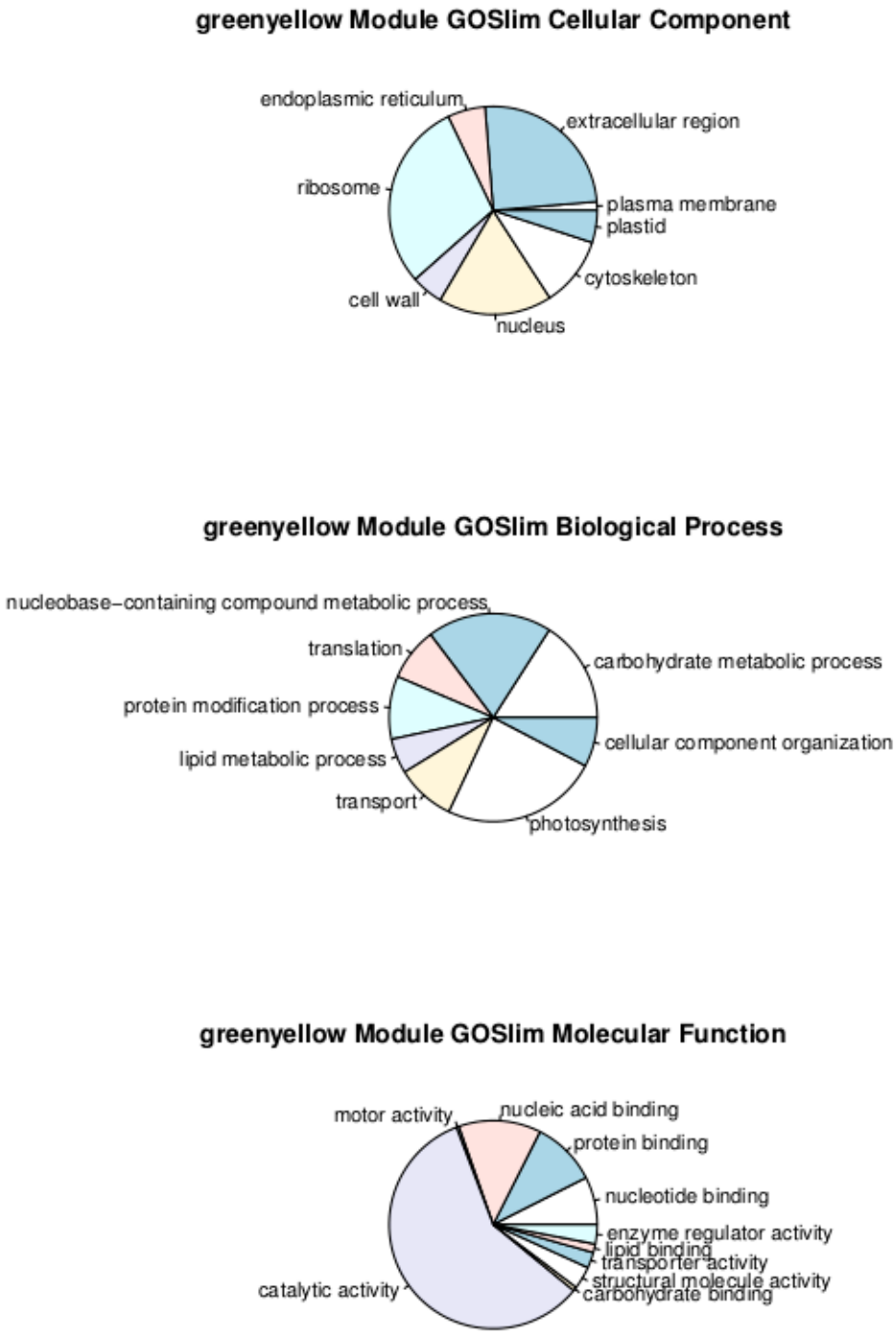


Figure A.13: Green Yellow module GOSlim classification distribution

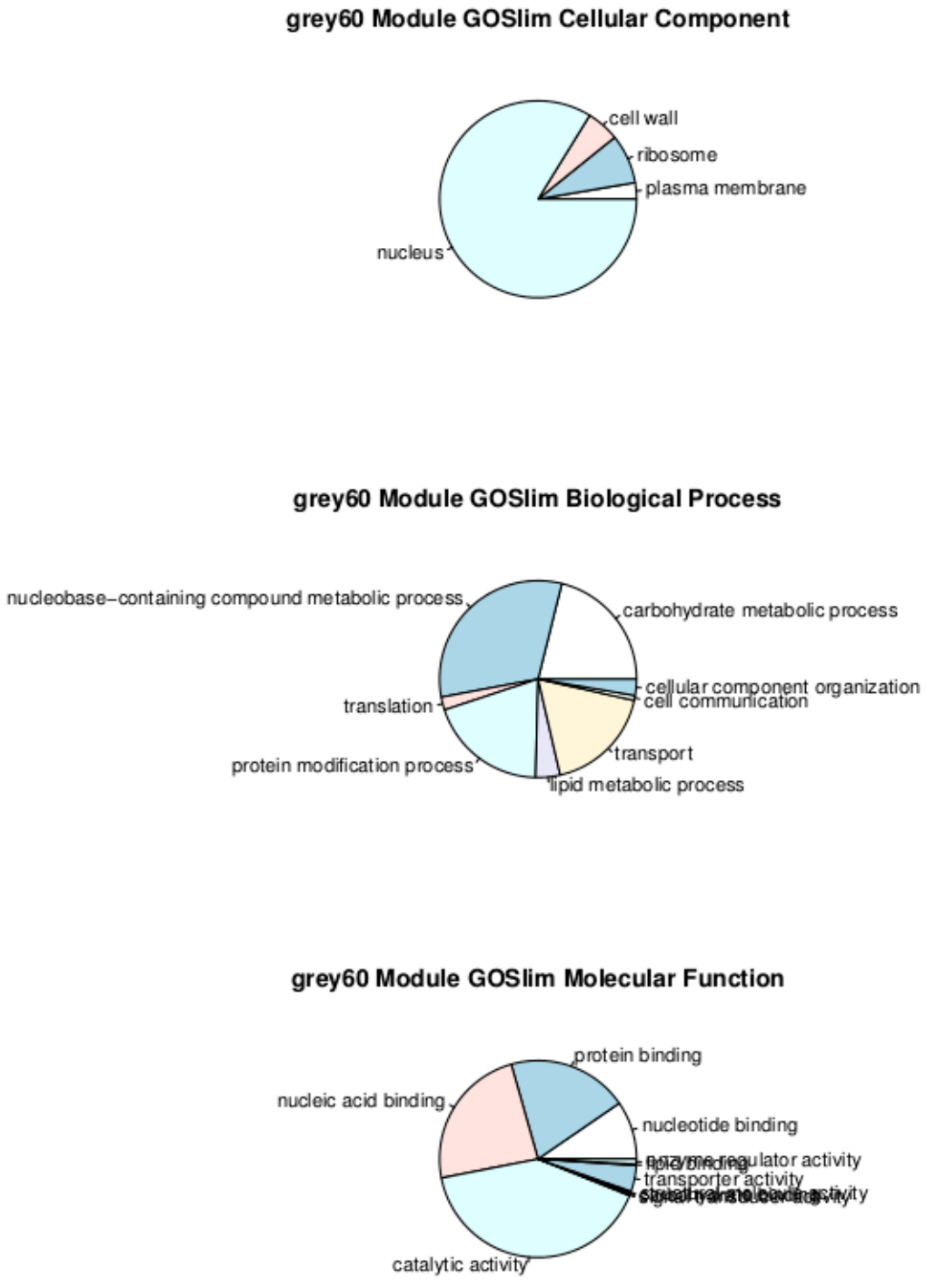


Figure A.14: Grey60 module GOSlim classification distribution

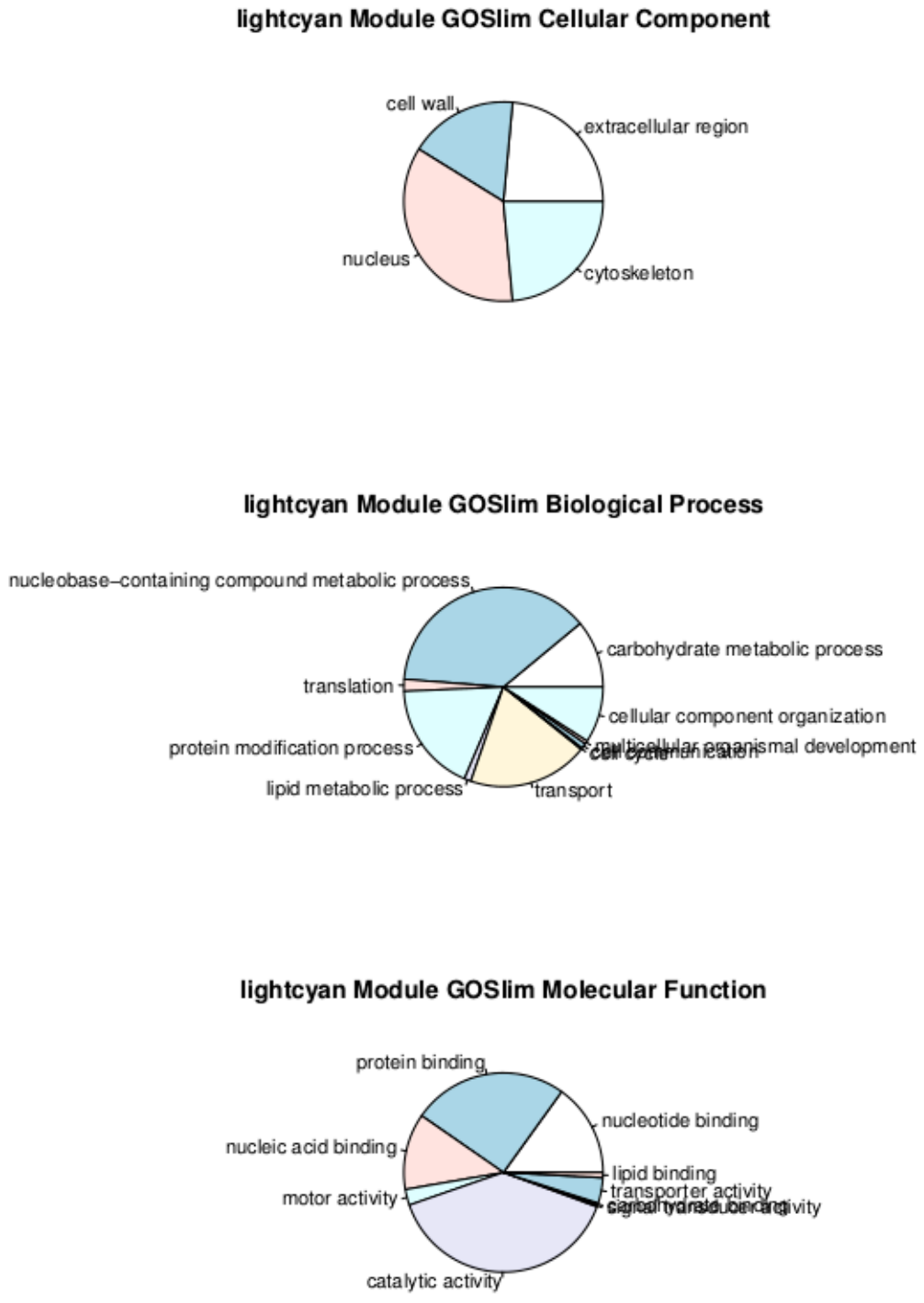


Figure A.15: Light Cyan module GOSlim classification distribution

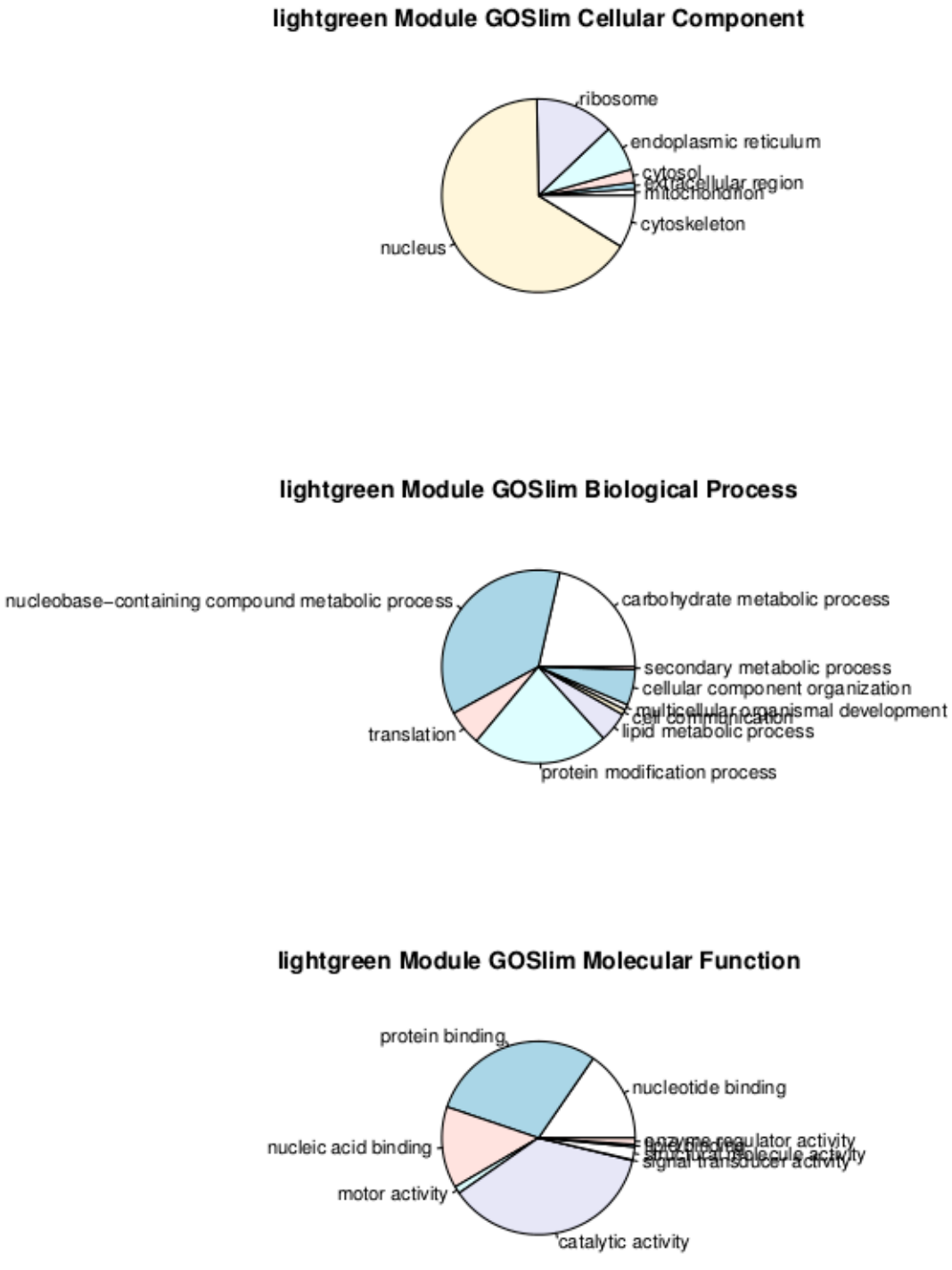


Figure A.16: Light Green module GOSlim classification distribution

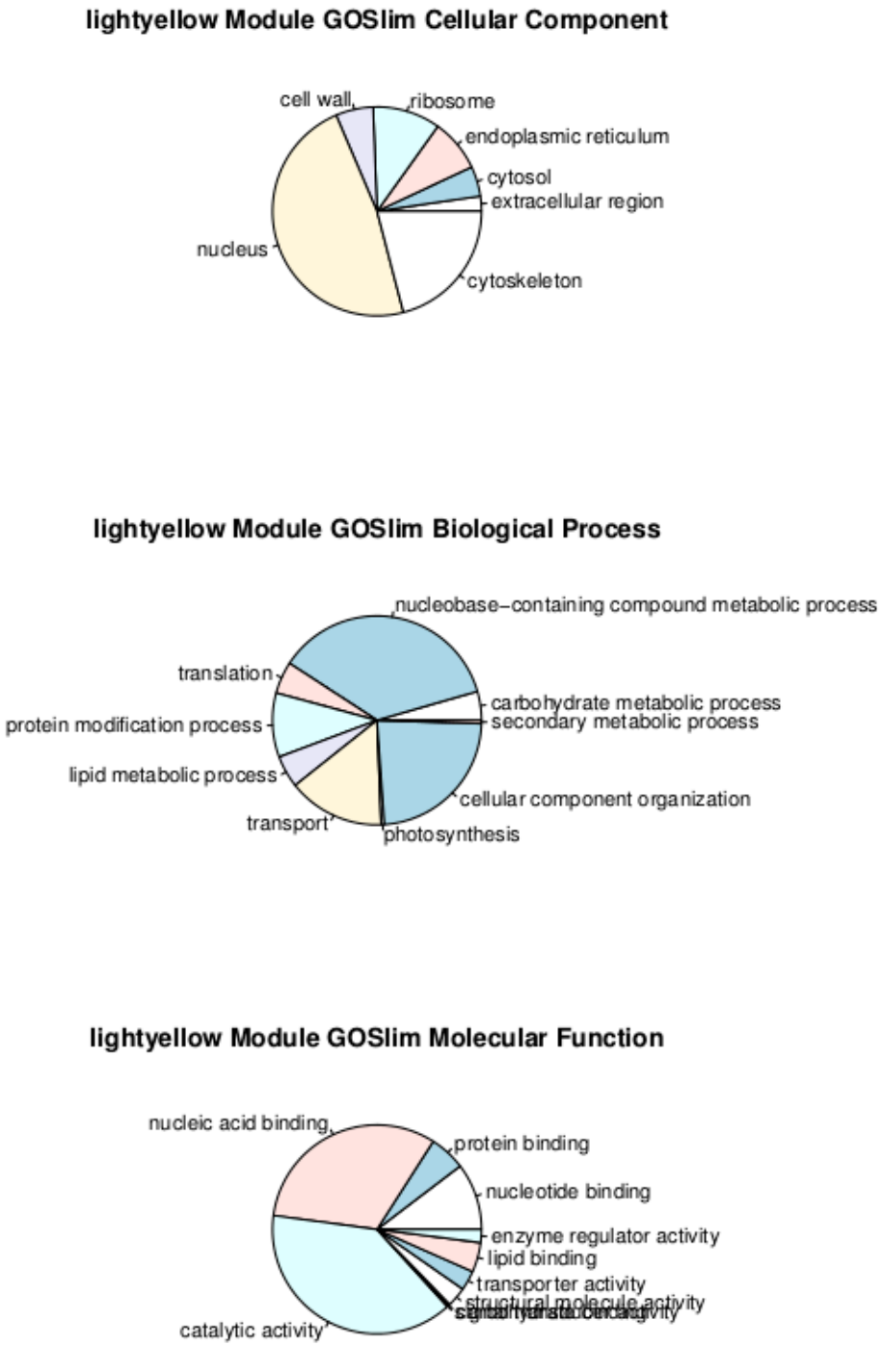


Figure A.17: Light Yellow module GOSlim classification distribution

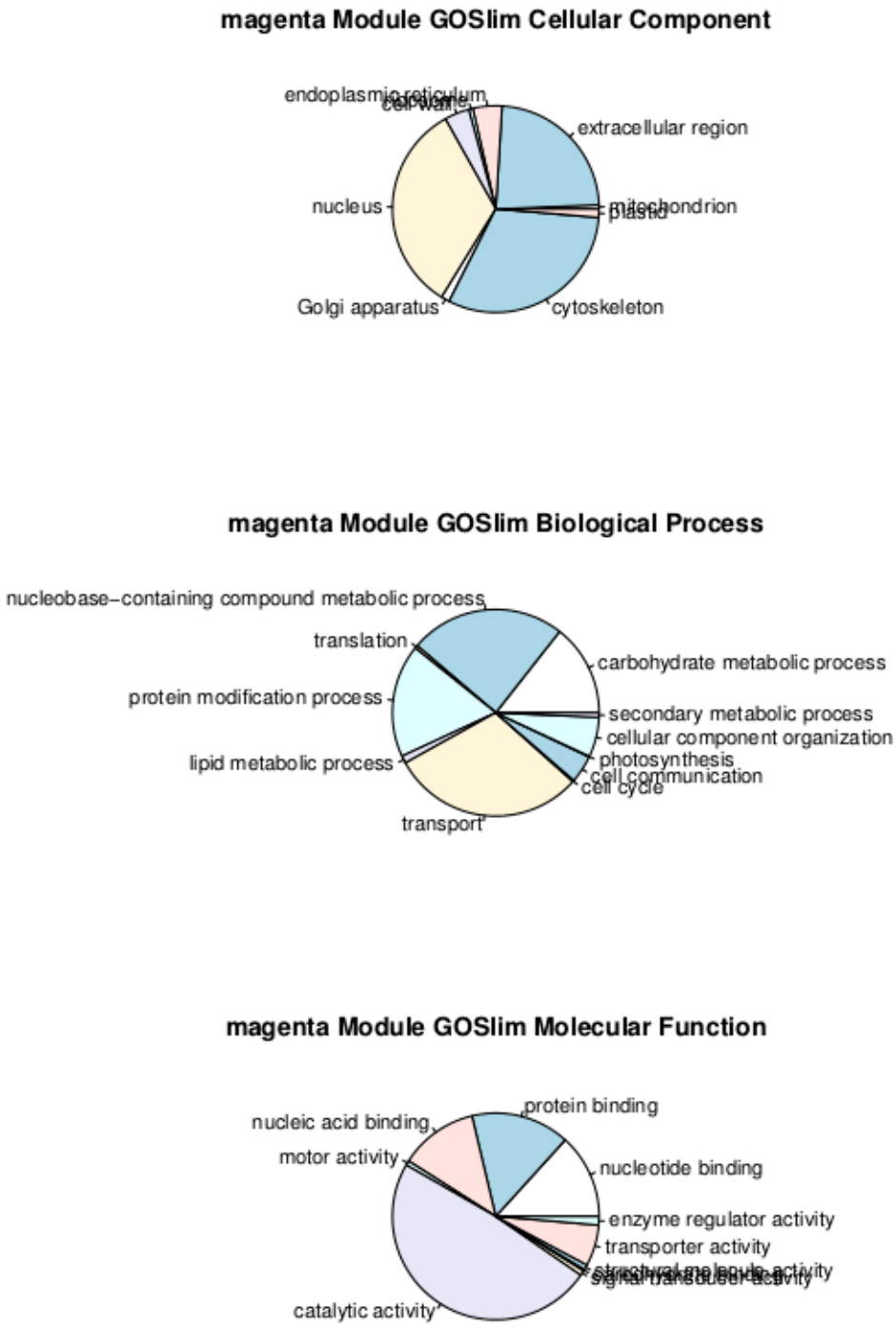
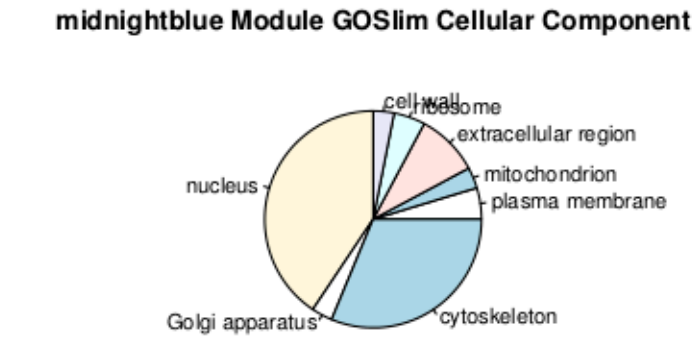
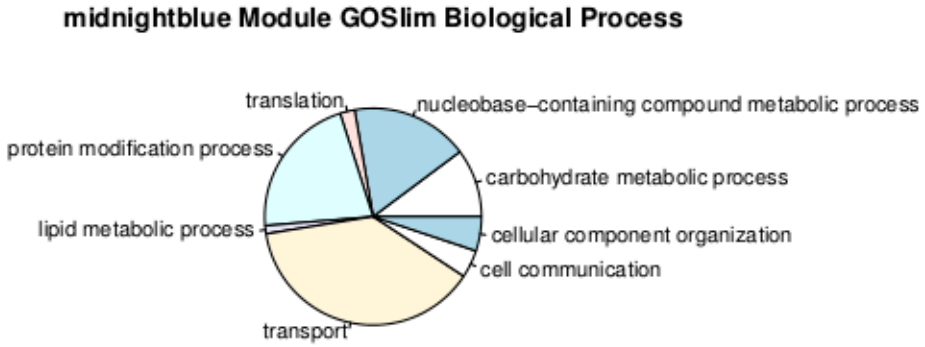


Figure A.18: Magenta module GOSlim classification distribution





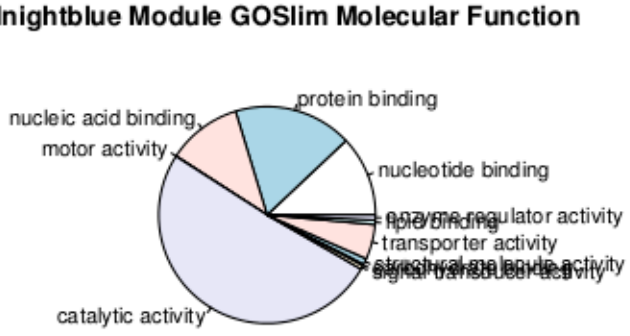


Figure A.19: Midnight Blue module GOSlim classification distribution

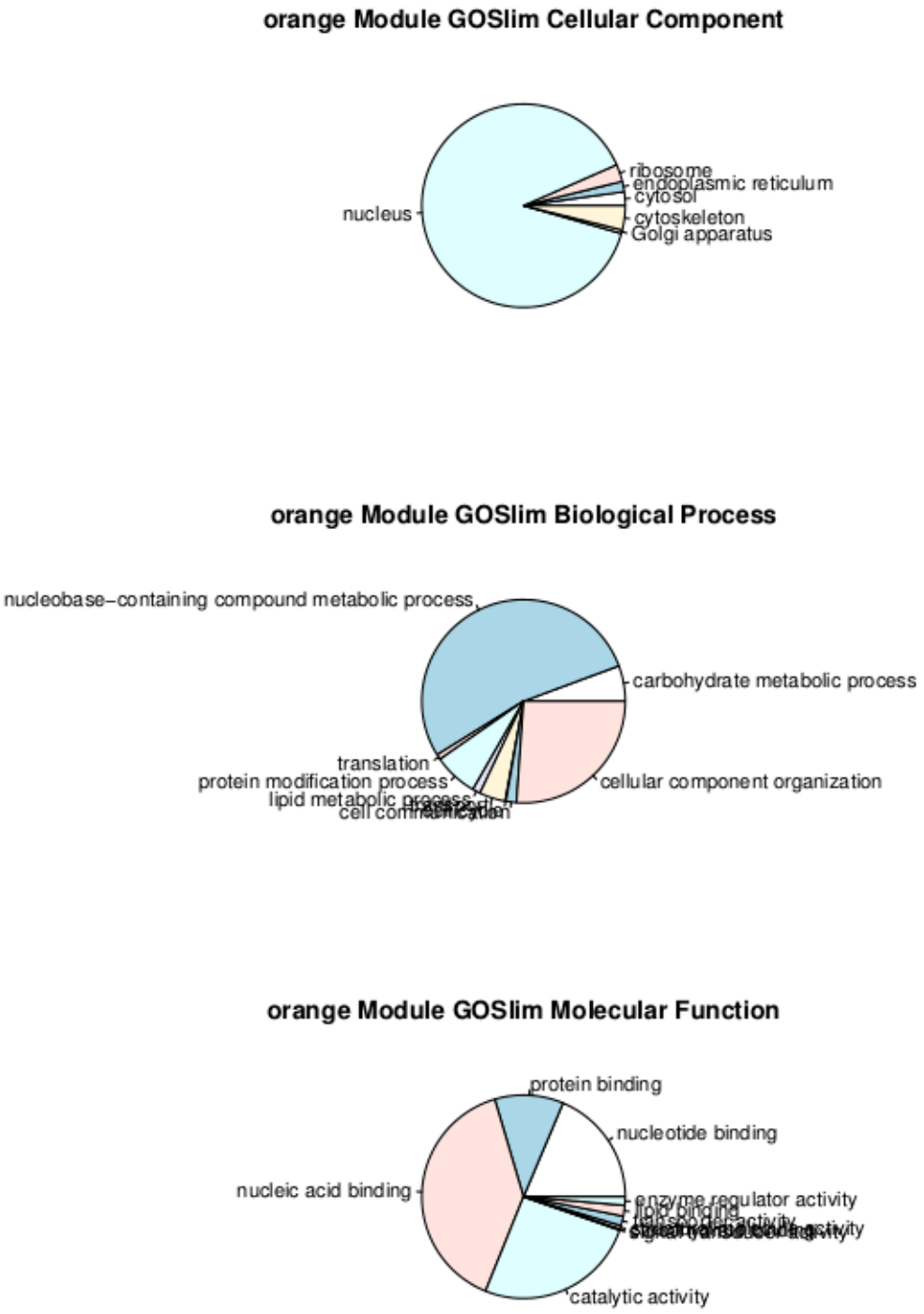


Figure A.20: Orange module GOSlim classification distribution

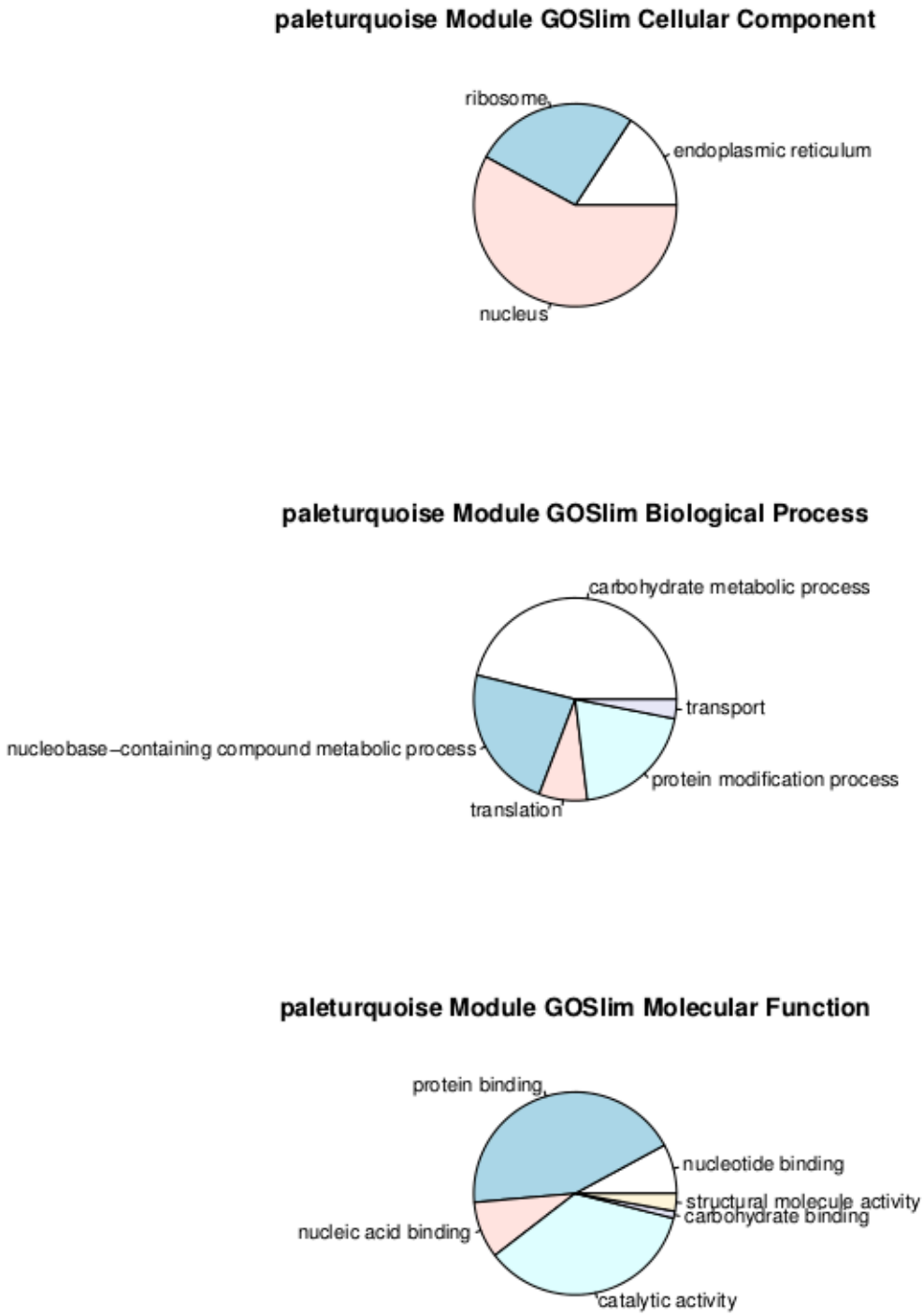


Figure A.21: Pale Turquoise module GOSlim classification distribution

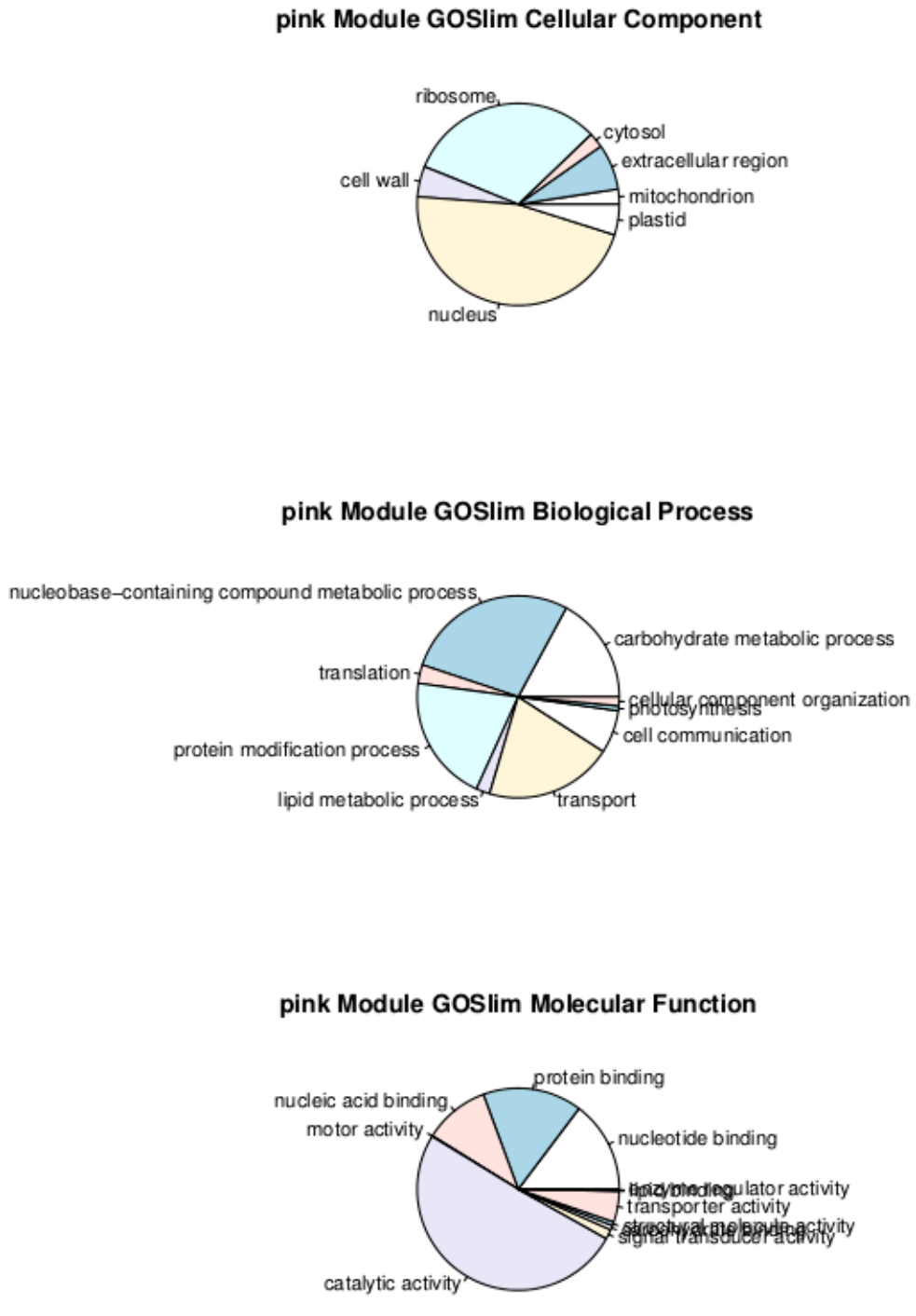


Figure A.22: Pink module GOSlim classification distribution

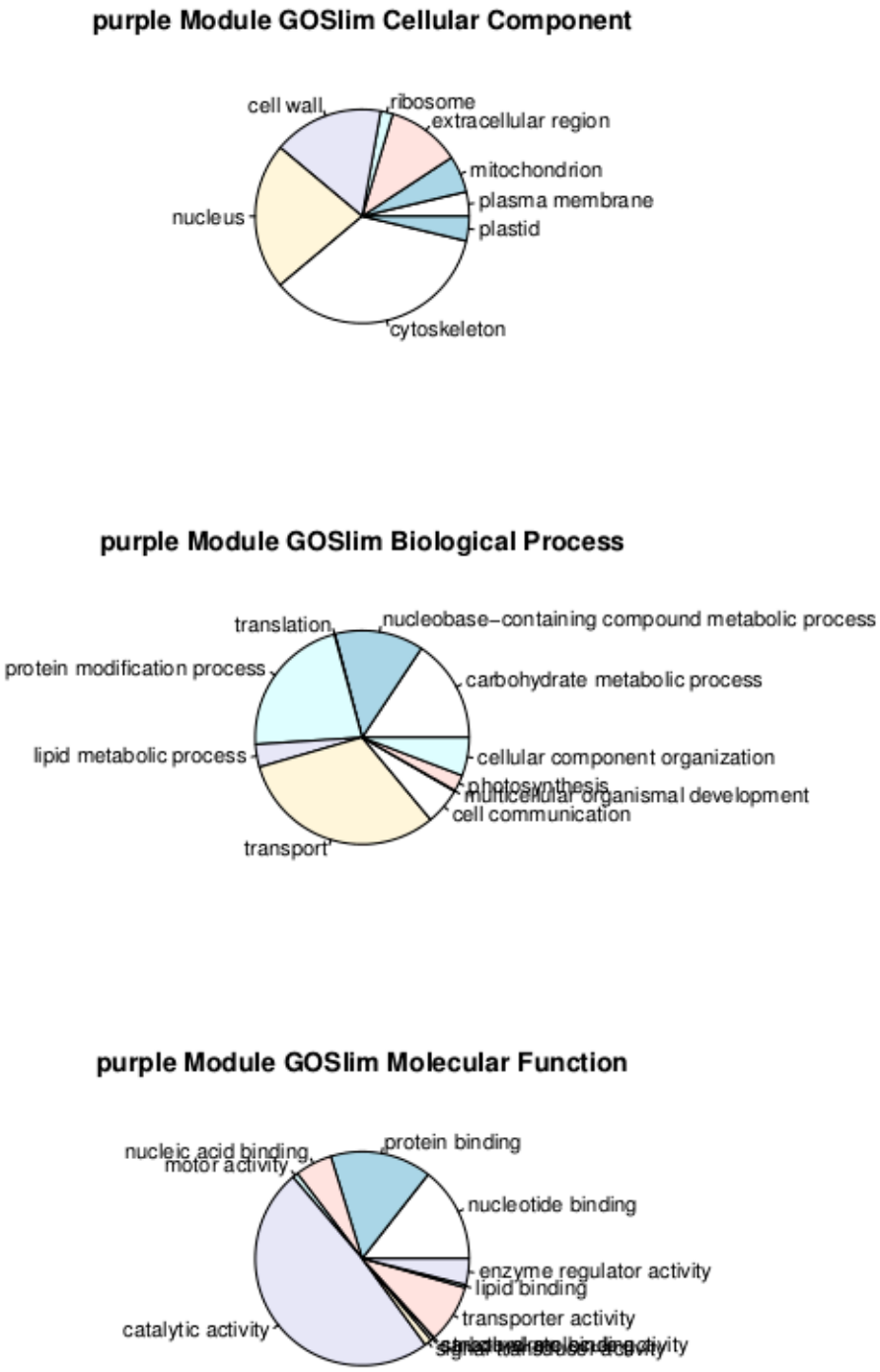


Figure A.23: Purple module GOSlim classification distribution

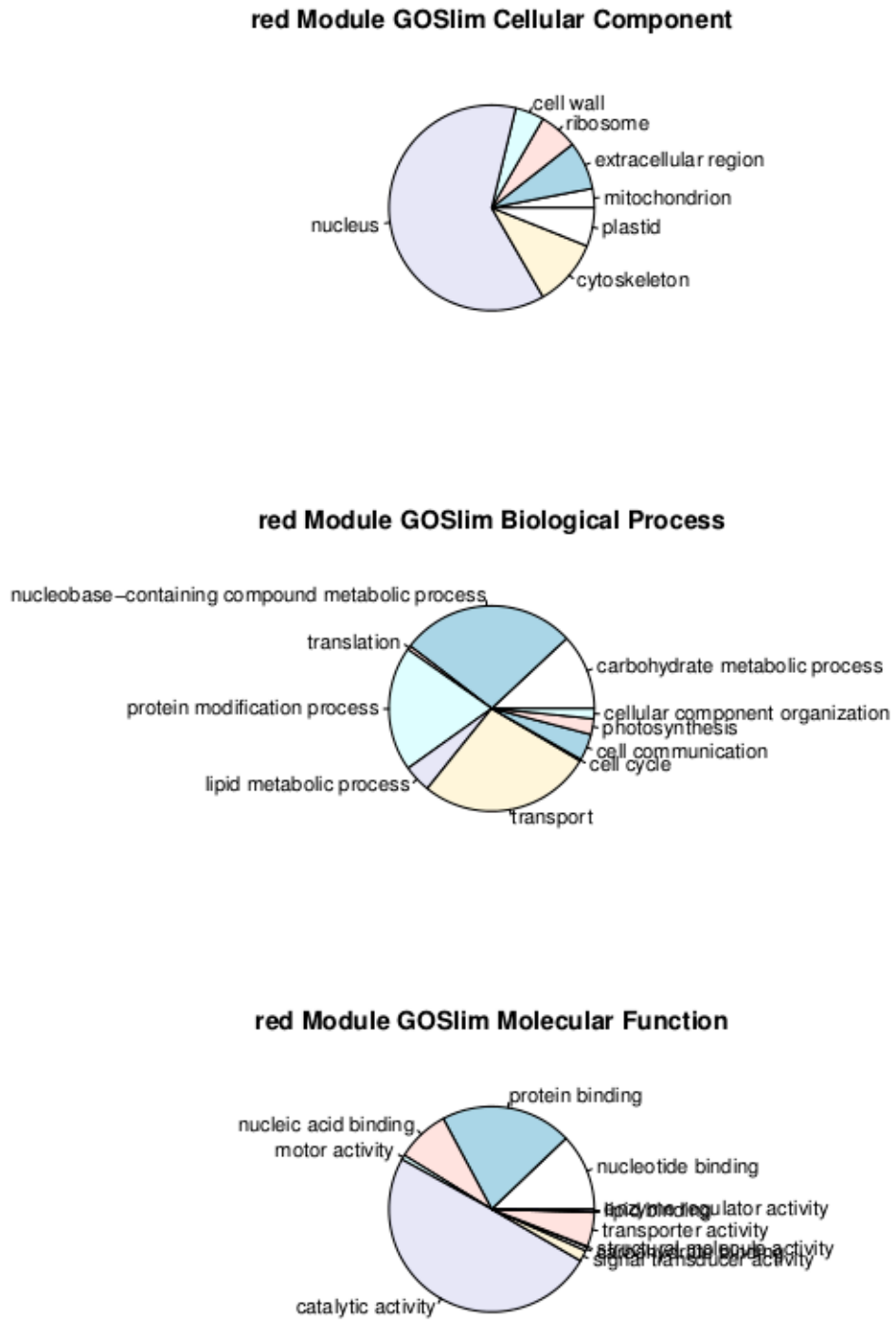


Figure A.24: Red module GOSlim classification distribution

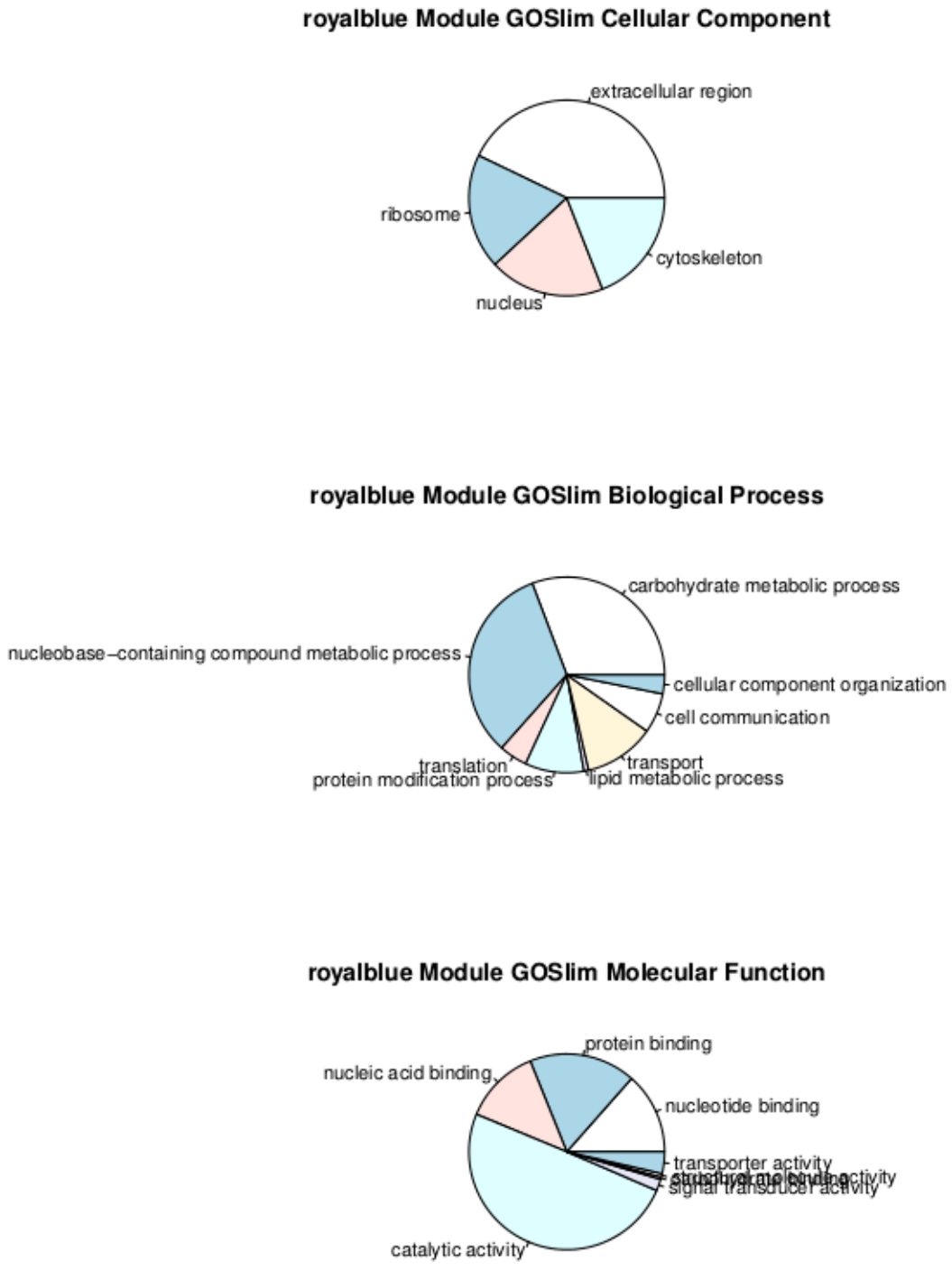


Figure A.25: Royal Blue module GOSlim classification distribution

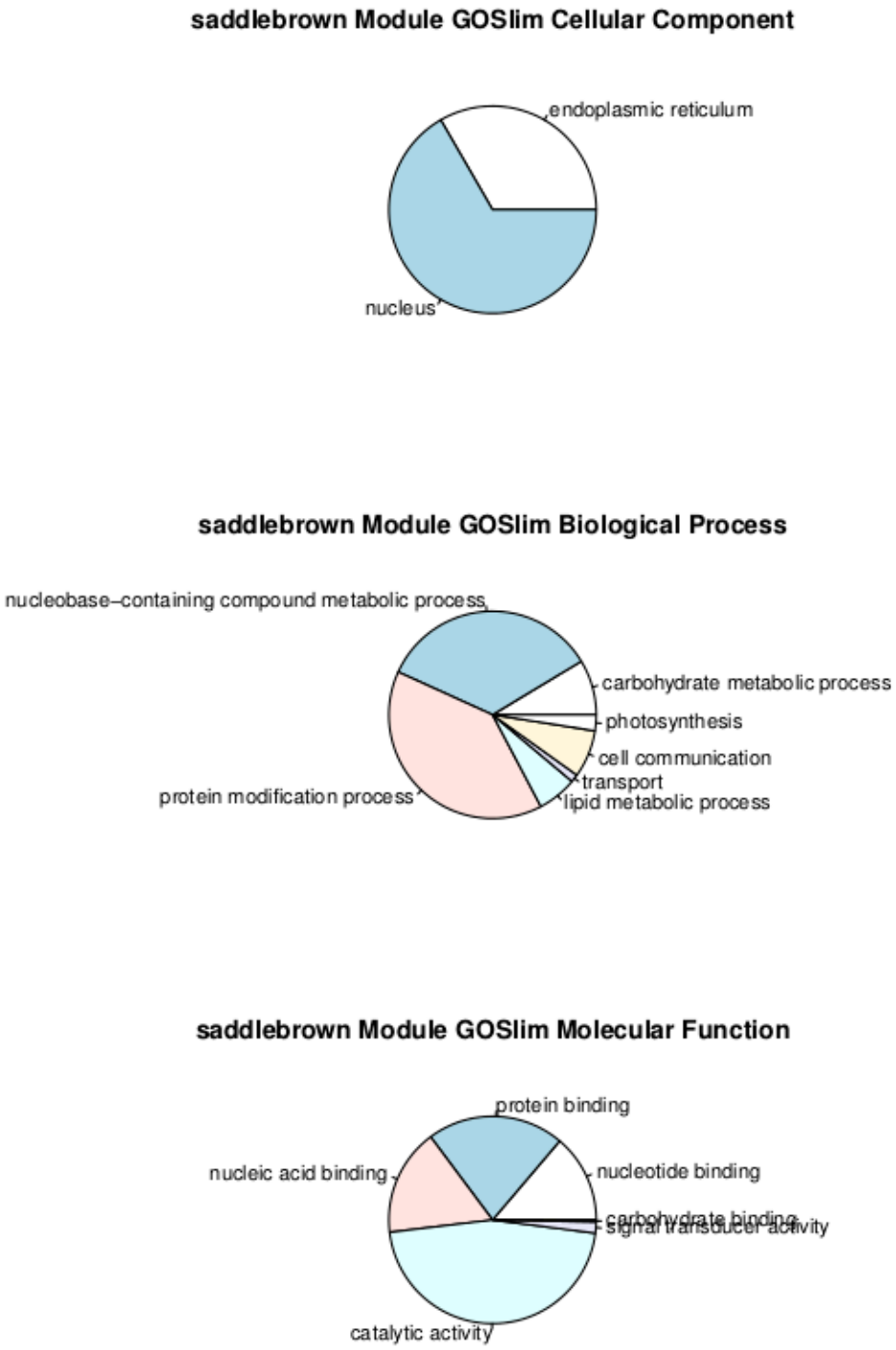
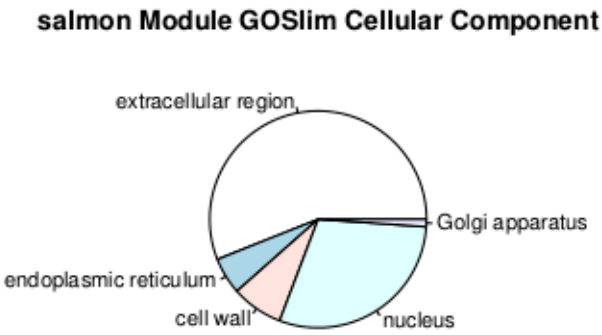
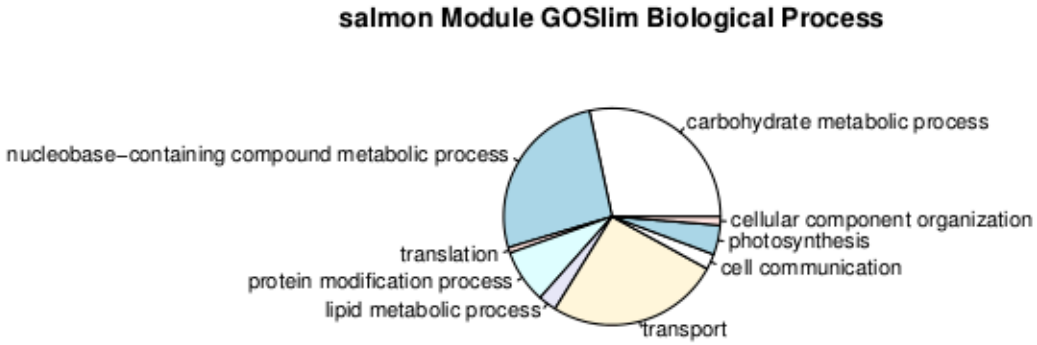


Figure A.26: Saddle Brown module GOSlim classification distribution





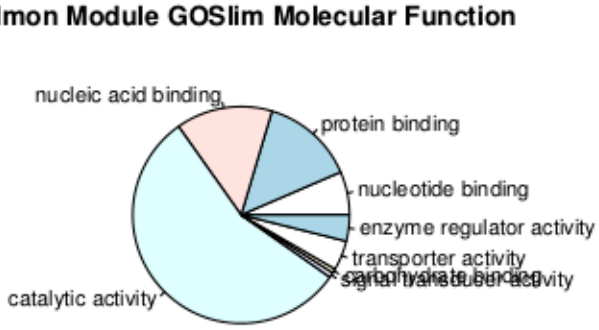


Figure A.27: Salmon module GOSlim classification distribution

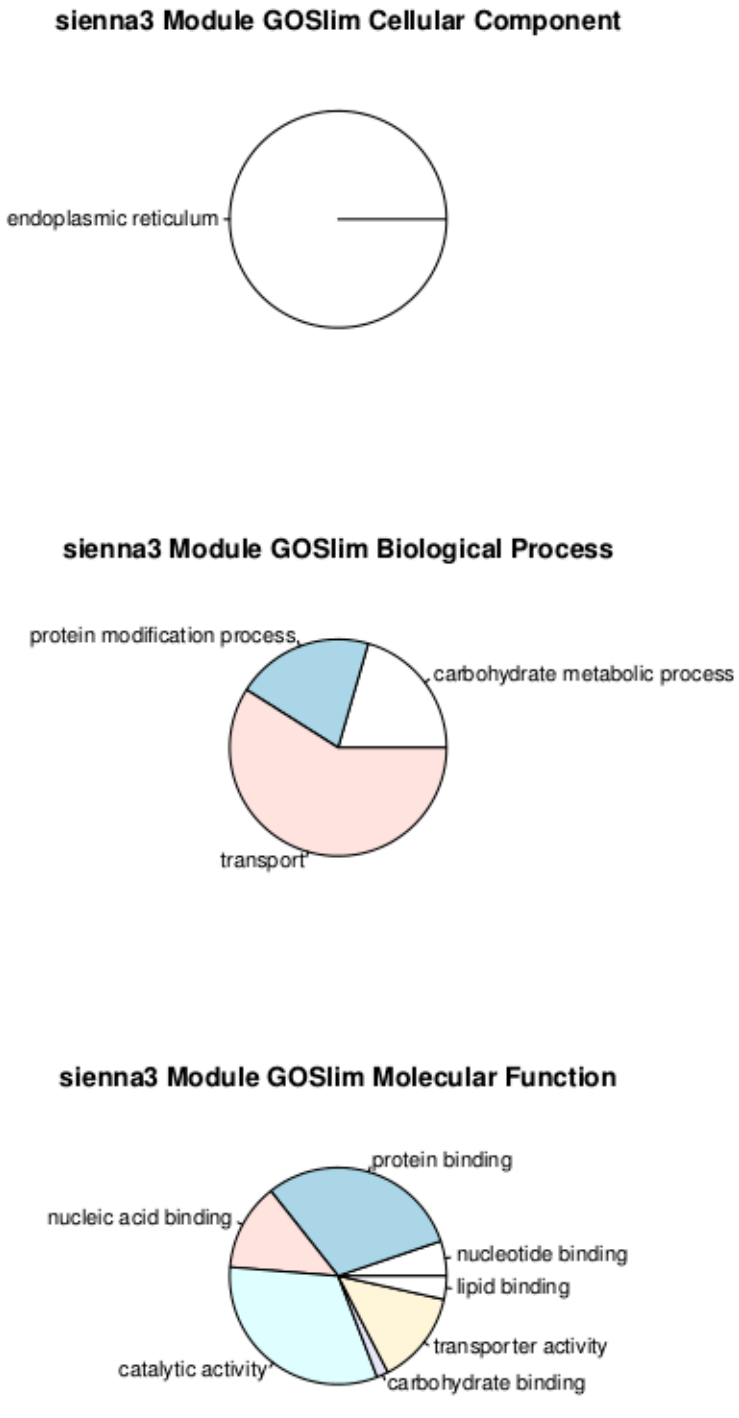


Figure A.28: Sienna3 module GOSlim classification distribution

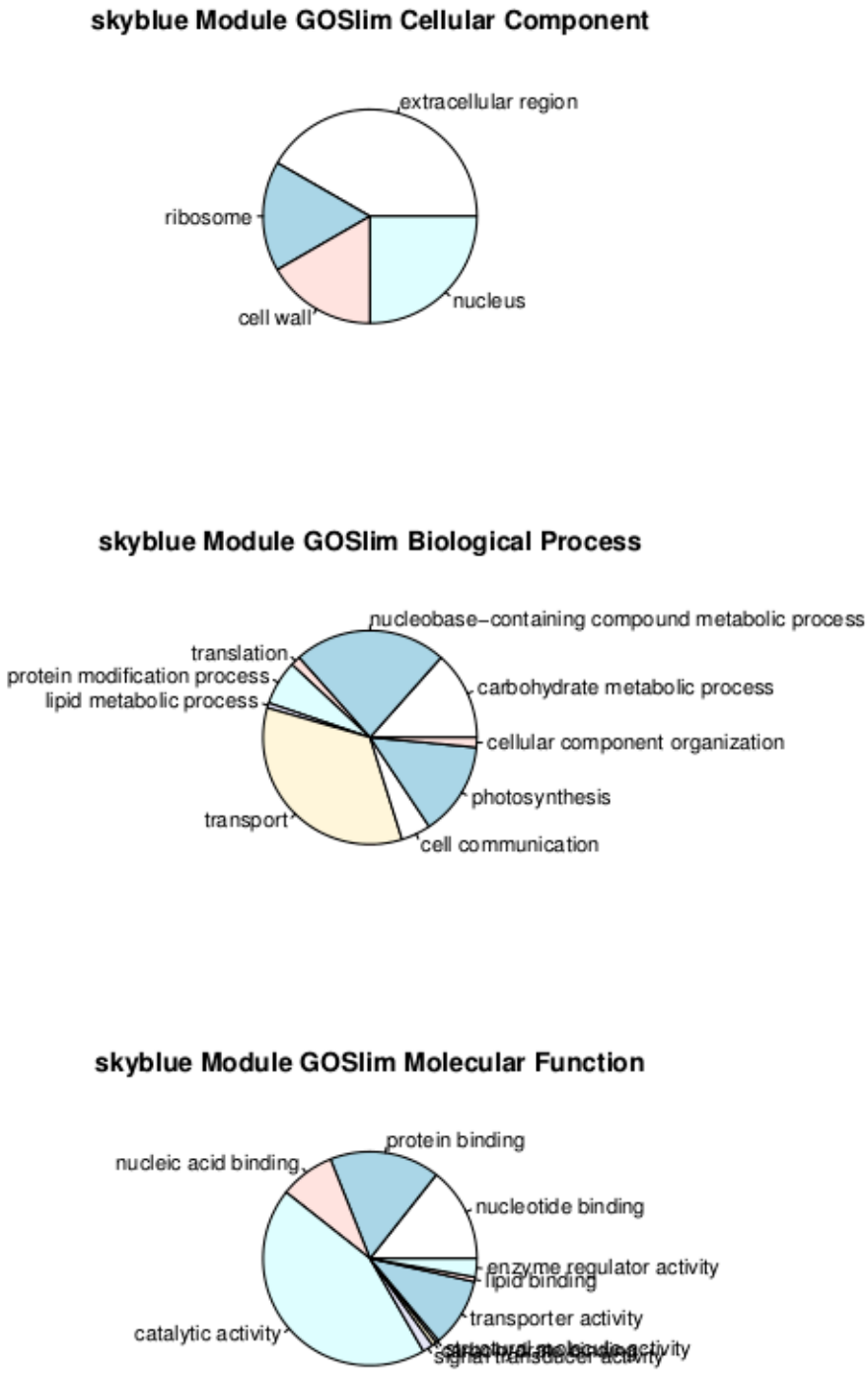


Figure A.29: SkyBlue module GOSlim classification distribution

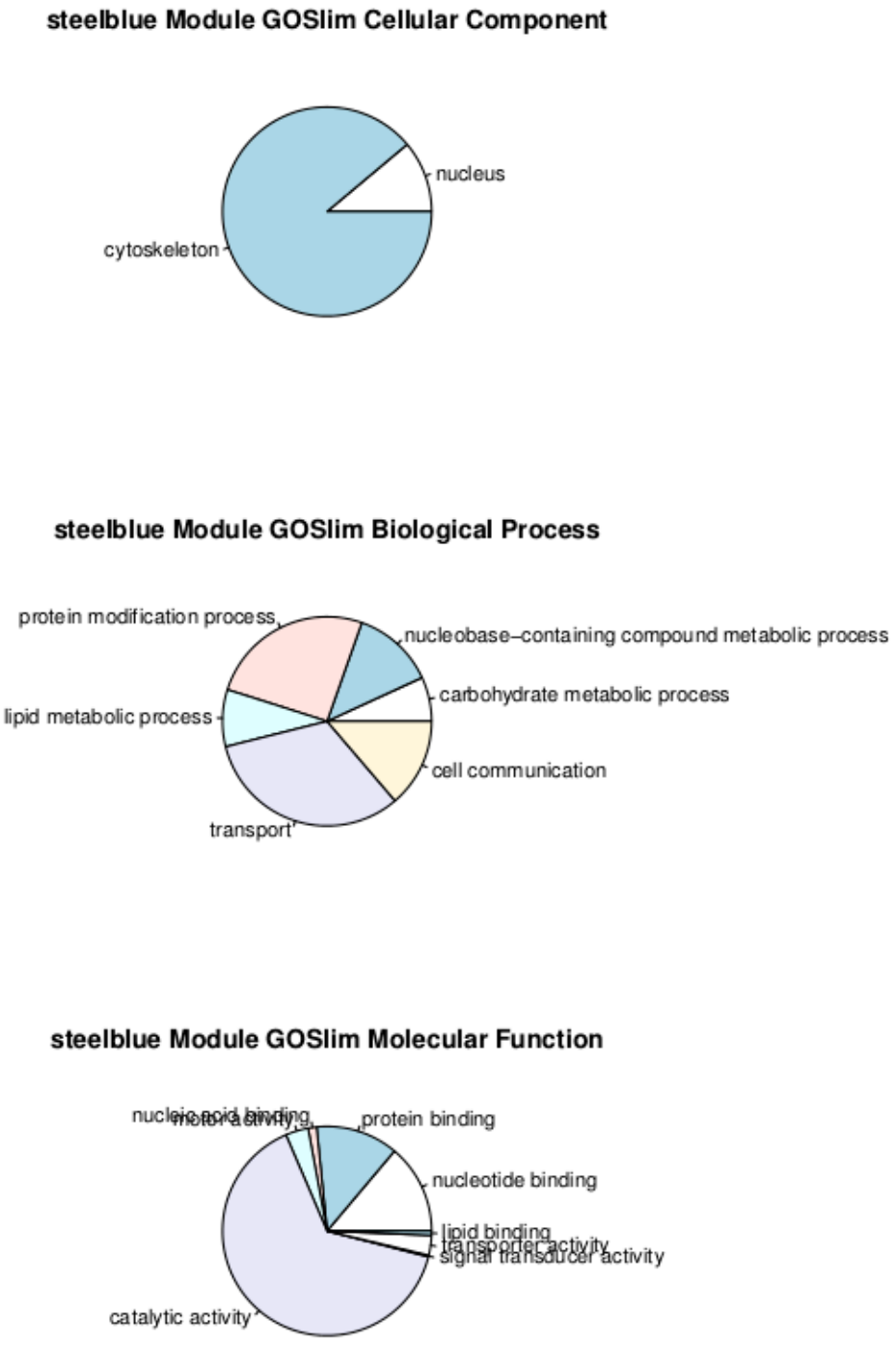


Figure A.30: Steel Blue module GOSlim classification distribution

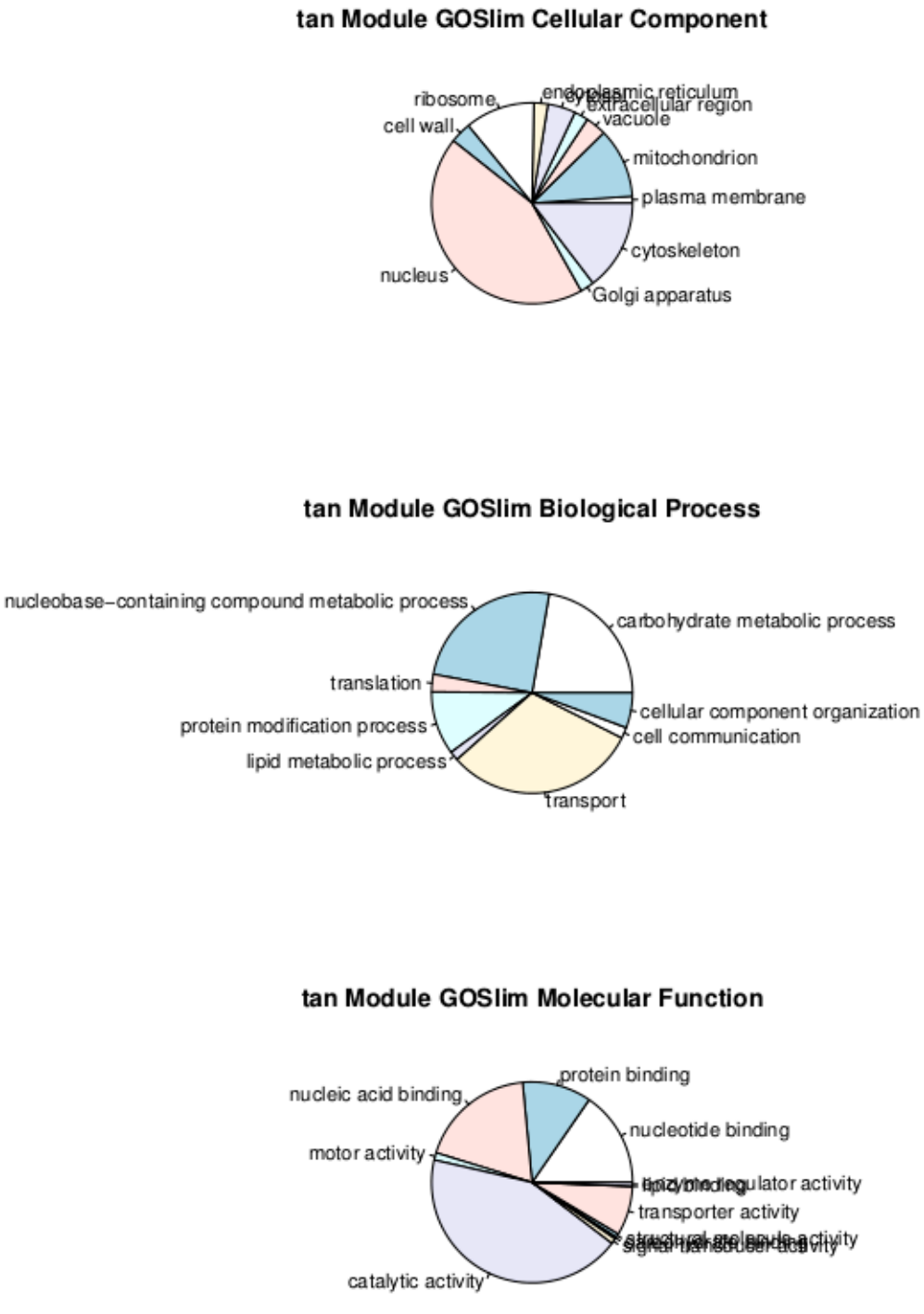


Figure A.31: Tan module GOSlim classification distribution

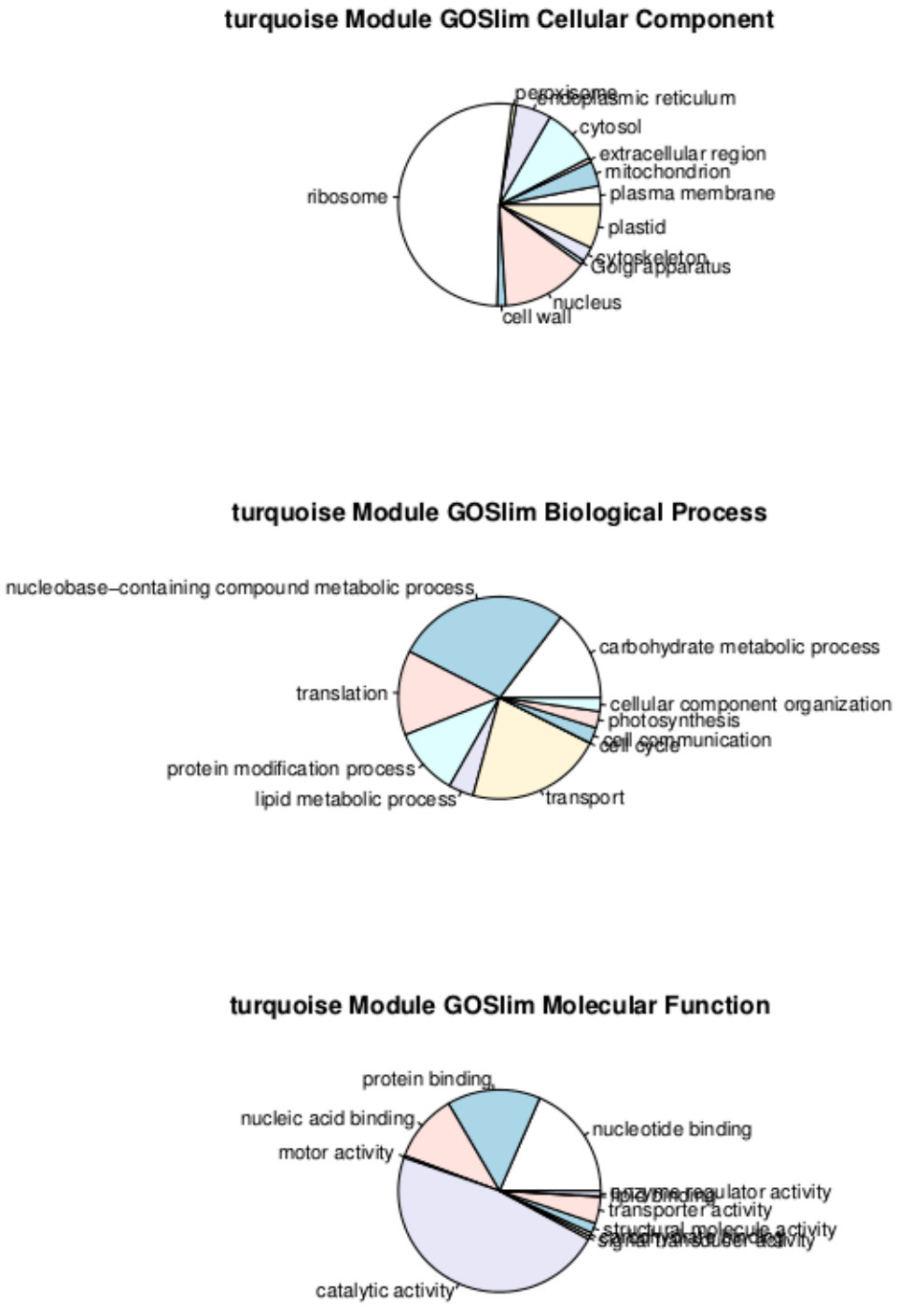


Figure A.32: Turquoise module GOSlim classification distribution

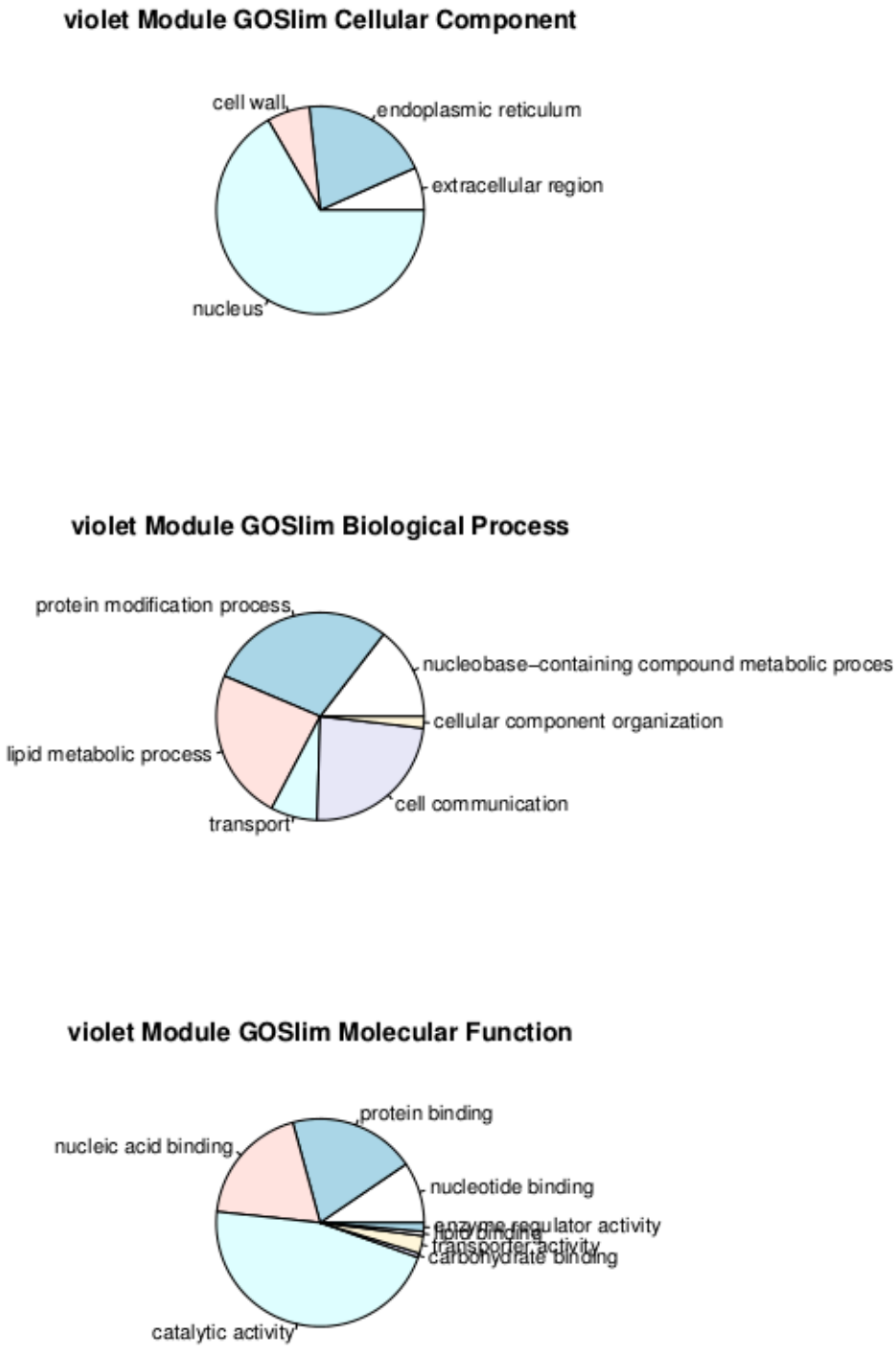


Figure A.33: Violet module GOSlim classification distribution

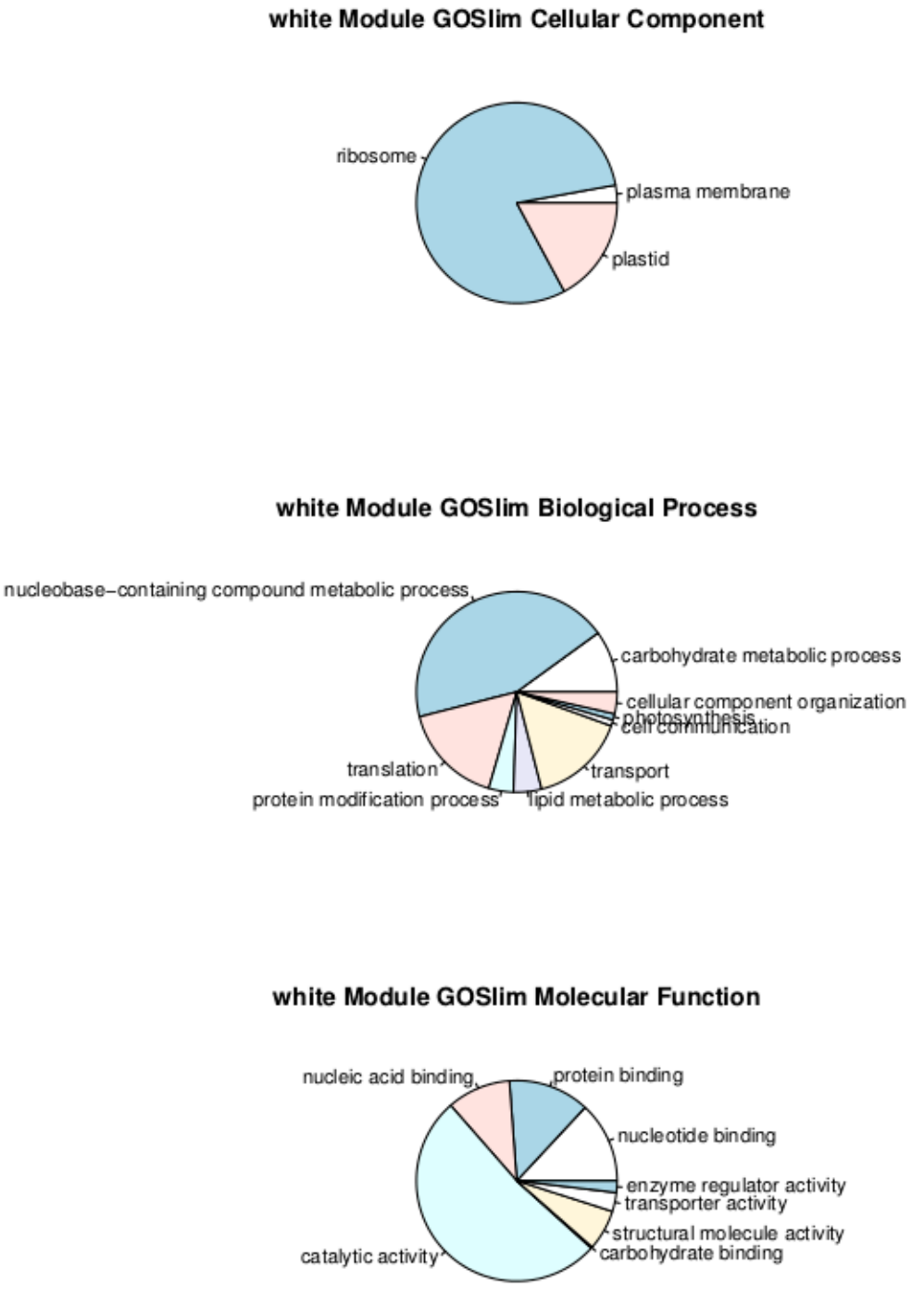


Figure A.34: White module GOSlim classification distribution

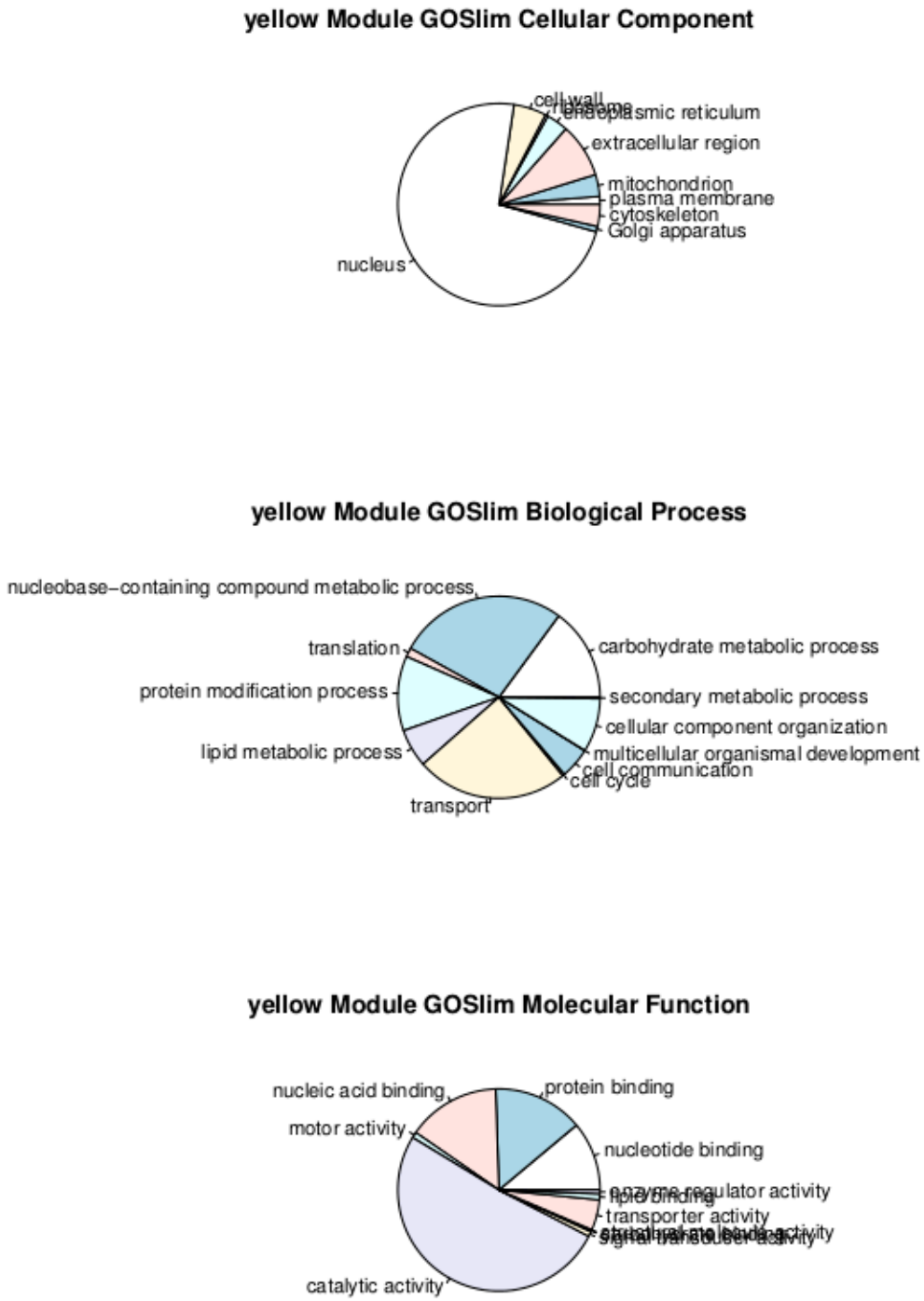


Figure A.35: Yellow module GOSlim classification distribution