Chapter 5: Using transcriptomics to investigate evolution and toxicology in *Gambierdiscus*. [1]

Key words: *Gambierdiscus*, ciguatoxin, pan-transcriptome

1 Abstract

The dinoflagellate species of the genus *Gambierdiscus* are reported to produce ciguatoxins (CTXs), the causative agent of ciguatera fish poisoning, a debilitating seafood-borne illness. Although the specific toxins produced by *Gambierdiscus* remain largely unknown, it has been verified using LC-MS/MS in multiple strains that the species *G. polynesiensis* produces CTX analogs. Bioassays have implicated several more species of *Gambierdiscus* with CTX-like bioactivity. An understanding for the evolution of *Gambierdiscus* and their toxins requires the availability of reference genomes. The fact, however, that *Gambierdiscus* species, as most dinoflagellates, possess very large (32-35 Gbps) and highly repetitive genome and complex genome architecture makes genome sequencing of these microalgae almost impossible. To overcome the challenges of genome sequencing, we generated *de novo* RNA assemblies of four species, *G. carpenteri*, *G. holmesii*, *G. lapillus* and *G. polynesiensis*, for a comparative study with the previously sequenced *G. australes* to determine transcripts that are shared amongst the investigated strains. Here we present a *Gambierdiscus* core transcriptome that might be used to investigate candidate genes related to toxin production.

2 Introduction

A challenging aspect of protist *de novo* sequencing projects lies in the lack of annotated closely related references for comparison, without which it is difficult to assess the adequacy and completeness of sequencing, library processing and assembly methods employed. This issue is particularly relevant for dinoflagellates, whose expansive and complex genomes and genetic machinery tend to be a barrier to genomic sequencing. As an alternative to wrangling with dinoflagellate genomes, transcriptomes can be used to explore their genetics. Dinoflagellates appear to have an uncharacterized genetic regulatory mechanism which leaves protein synthesis regulation to the post-transcriptional stage. For this reason, mRNA may give an approximation of genomic content. An indication of these regulatory mechanisms comes from a number of direct previous observations. Harke et al. (2017) cultured the dinoflagellates *Prorocentrum minimum* and *Alexandrium monilatum* under stress conditions by severely limiting nitrogen as well as phosphorous availability. The cultures showed significant biochemical changes (e.g. altered growth rate, particulate organic carbon and particulate carbohydrates content) between the control and stress conditions at time of harvest, yet change in transcriptome expression was minimal, between 0.1 to 1 % depending on stressor and species used [10]. While the difference in biochemical changes was not captured by mRNA profiling of the cultures, the study did not include a protein expression observation to verify a difference in expression despite a static pool of mRNA availability [10].

Due to the relative difficulty in culturing dinoflagellates and extracting their RNA, the number of marine eukaryotic transcriptomes was limited until the Mmrine microbial eukaryote transcriptome sequencing project (MMETSP) data release. When searching for *Gambierdiscus* on NCBI's SRA database, 5 relevant projects were found in addition to the MMETSP results (searched on November 10, 2018). These sequencing projects covered two strains of *G. polynesiensis*, as well as *for G. australes* and *G. excentricus*. The fifth project focused on the bacterial associations of *G. caribaeus* and *G. carolinianus*. Broadening the search to the order gonyaulacales yielded a further 19 projects, including another on bacterial associates as well as 3 projects on *Azadinium* and *Crypthecodinium*, which are arguably not part of the gonyaulacales (see **chapter 4**). Searching for members of the phylum dinoflagellates yields a further 84 projects. Despite their ecological relevance for nutrient cycling, dimethylsulfoniopropionate production, coral symbiosis

and neurotoxin production (for a review see [26]), a paucity of sequencing data, even with the MMETSP dataset, is evident. This is further confounded by a large proportion of dinoflagellate transcripts sharing little or no similarity to known proteins or domains in public databases. When compared to NCBI's nr database, the proportion of contigs with no known match was 60 % for *Azadinium spinosum* [24], over 50 % for *G. australes* & *G. belizeanus* [18], 57.9 % for *G. excentricus* [17], 63 % for *G. polynesiensis* [17, 27], and 55 - 57 % for *Karenia brevis* [34].

The concept of a reference genome or transcriptome allows for direct comparison of sequencing data to a high quality standard. However sequencing further genomes in bacteria revealed a large transitory subset of genetic content, with the conclusion that a single strain-based reference would be inadequate for capturing a large proportion of the species' genetic diversity [36, 37]. An alternative approach to defining a genomic reference for a species was proposed: that of pan-transcriptomics, where a core genome common to all strains, and an accessory genome which is strain specific. Since then the pan-genome, or transcriptome, concept has been adopted for eukaryotes also, with the realization that the accessory genomic content exists when multiple strains of a species are sequenced (e.g. [14, 23, 28, 29, 31, 35]). The pan-genome analyses concept has been applied at different taxonomic levels, from intra-genus to super kingdom [11, 14, 19, 21, 37].

Five transcriptomes of *Gambierdiscus* were compared in this study with the aim of providing a pan-transcriptome for *Gambierdiscus de novo* transcriptome sequencing, which can be expanded and refined in future studies. The taxa originated from two locations in Australia (Merimbula, NSW, and Heron Island, QLD) and Rarotonga in the Cook Islands (Table 1). All of the five species have shown MTX-like bioactivity, while *G. carpenteri* did not register for CTX-like activity in a bioassay [22]). The toxin profiles registered all species apart from *G. carpenteri* as MTX producers, while only *G. polynesiensis* had a confirmed CTX production profile (Table 1). This study revealed a set of core-transcripts shared by all taxa as well as a subset of species-specific, accessory portions of the transcriptome. The results in this study could provide an avenue of investigation of querying the expression differences between toxic and non-toxic species of *Gambierdiscus*.

Table 1: *Gambierdiscus* species transcriptomes used in this study along with their toxicity, toxin profile, accession numbers and source. Where possible, information is strain specific & otherwise denoted with *

Species	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G. holmesii*
Strain	CAWD149	UTSMER9A	HG4	CG15	HG5
Transcriptome source	MMETSP	**chapter 4**	**chapter 4**	**chapter 4**	**chapter 4**
Accession ID	MMETSP0766	SRR6821720	SRR6821722	SRR6821723	SRR6821721
Isolation location and year	Rarotonga, Cook Islands (2007)	Merimbula, Australia (2014)	Heron Island, Australia (2014)	Rarotonga, Cook Islands (2014)	Heron Island, Australia (2014)
Toxin profile (LC-MS/MS)	CTX -ve; MTX +ve	CTX -ve; MTX -ve	CTX -ve; MTX +ve	CTX +ve; MTX +ve	CTX -ve; MTX +ve
Toxicity via bioassay	CTX +ve; MTX N/A	CTX -ve; MTX +ve	CTX +ve*; MTX +ve*	CTX +ve*; MTX +ve*	CTX +ve*; MTX +ve*
References	[16, 25, 32]	[22]	[20, 22]	**this study**, [4]	[20, 22]

3 Methods

Scripts used for this project are available on Github under hydrahamster/pan-tran. Venn diagrams were created with InteractiVenn [12]. Bar graphs were generated in Origin Pro (OriginLab, Northampton, MA).

3.1 Transcriptome acquisition

Species of *Gambierdiscus* used in this chapter are summarized in Table 1. Unless otherwise noted, toxicity and toxin profile reports are specific to the strains used as inter-species variation in toxin production has been recently reported [22, 33]. The *G. polynesiensis* toxin profile was elucidated by Tim Harwood at the Cawthron Institute (Nelson, NZ) with the same methodology as for *G. lapillus* used in **Chapter 2**. RNA-seq libraries were assembled as per the transcriptome assembly subsection in the methods of **Chapter 4**, without diginorm.

3.2 Spliced leader search

The dinoflagellate spliced leader (dinoSL) sequences reported by Zhang et al. (2007) were used to build a HMMER library [39]. The transcriptome assemblies were searched with the dinoSL hmmer library to investigate for spliced leader presence. All clusters were searched for membership of one or more contigs with a dinoSL.

3.3 Homolog clustering

Cd-hit was used to cluster highly similar transcripts to reduce redundancy with the flags -T 10 -M 5000 -G 0 -c 1.00 -aS 1.00 -aL 0.005 as shown by Cerveau and Jackson (2016) [3, 7]. Transdecoder was use to predict coding regions on the clustered nucleotide sequences [9]. Protein clusters were annotated with Interproscan v5.27 with local lookup server [30]. Protein clusters were processed to include the species of origin instead of the TRINITY tag and concatenated for input to get_homologues [38]. The -t 0 flag was used for get_homologues to acquire all possible clusters even with only one species representative, and -G for the OMCL algorithm. The core transcriptome was defined as

clusters with members from all five species. Clusters containing four of the five species were considered as the softcore. The accessory transcriptome consisted of clusters from a single species. The resulting core, softcore and accessory clusters were matched with their interpro annotations and GO terms were queried with GOSUM against the basic Gene Ontology (GO) database [1, 5, 15]. GOSUM was run at levels 1 and 2 of GOs with the go-basic GO reference.

3.4 Ketosynthase domain search

The transcriptome assemblies were queried for the ketosynthase (KS) active domain of the polyketide synthase (PKS) enzyme using HMMER [6] with libraries developed for this project. The contigs which were identified to contain an active domain were then searched for within the clusters to identify how the active domains clustered; and the assemblies were searched to compare KS abundance between species.

4 Results

4.1 Overview of the transcriptomes

The progression of clustering and annotation results per transcriptome can be found in Table 2. A total of 287,546 clusters were found across all five species (Fig. 1).

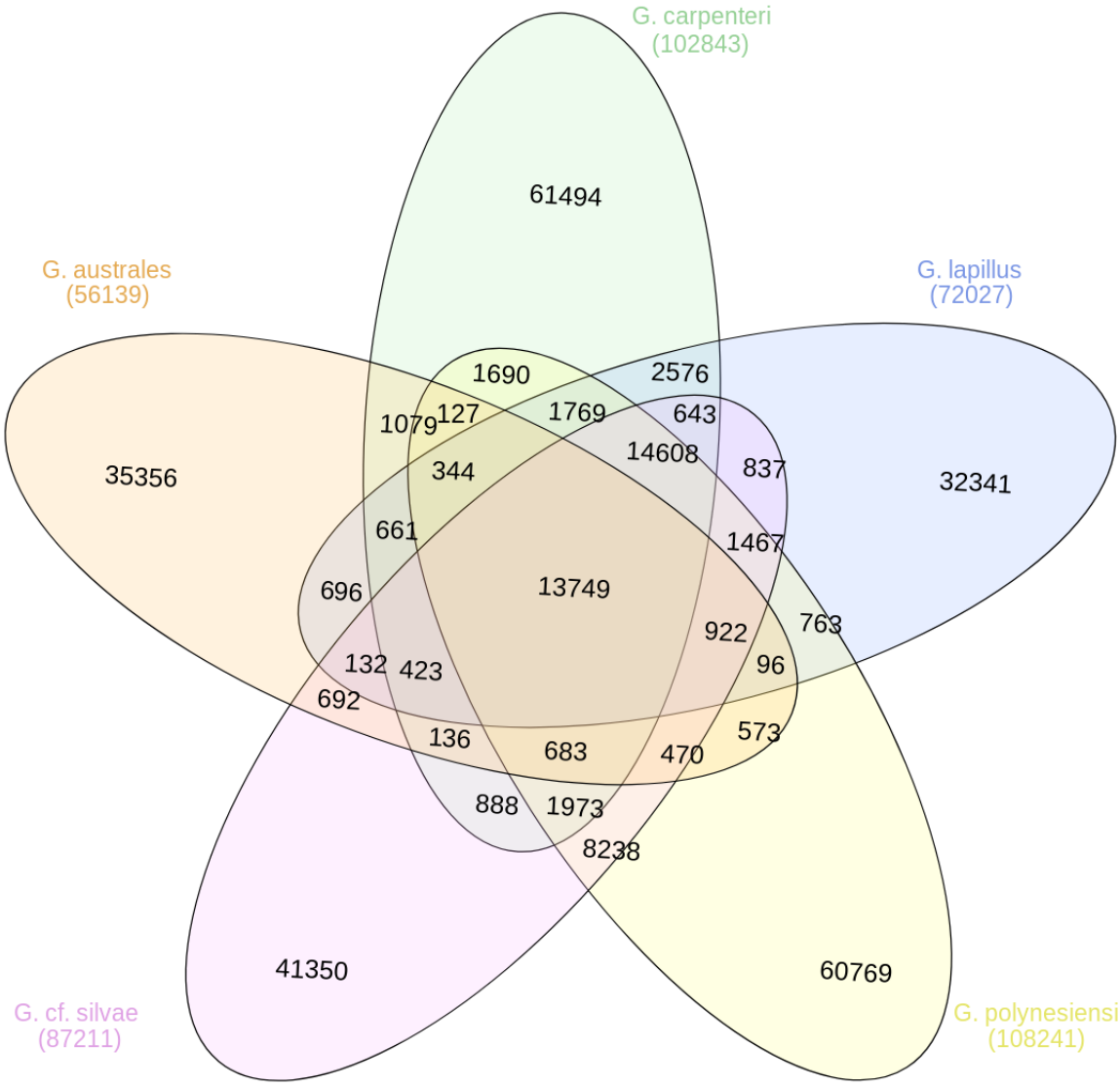

Figure 1: Venn diagram of cluster distribution between the five *Gambierdiscus* species. Core, accessory and inter-species cluster cross over depicted.

Table 2: Progression of clusters found in each *Gambierdiscus* transcriptome during processing.

Species	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G. holmesii*
Contigs	102,863	263,829	148,972	270,315	191,224
Spliced leader contigs	304	683	232	1,570	1,524
Nucleotide clusters (cd-hit)	102,861	263,743	148,966	270,265	191,205
Predicted coding regions (Transdecoder)	63,299	180,568	111,862	176,290	132,688
Contigs annotated (Interpro Scan)	131,970	334,737	225,324	225,324	254,844
Core-transcriptome clusters	13,750	13,750	13,750	13,750	13,750
Softcore-transcriptome clusters	2,372	16,058	16,297	16,557	16,636
Unique clusters	35,356	61,494	32,341	60,769	41,350

4.1.1 Core transcritome

A set of core genes common to all five species of *Gambierdiscus* were found. This set consisted of 13,750 amino acid clusters (Table 2) of which 45 % were annotated with GO terms (Suppl. table **??** & **??**). The largest core cluster contained 180 contigs with 23, 45, 32, 31 and 49 from *G. australes*, *G. carpenteri G. lapillus*, *G. polynesiensis* and *G. holmesii*, respectively, and was of unknown function. Twelve of the core clusters contained 100 or more contigs, of which 3 were unannotated. The predicted protein coding regions for the other nine clusters, in descending order of contig numbers are: an enzyme with catalytic activity involved in metabolic process, a calcium binding transmembrane transport channel, a protein involved in calcium binding, a protein binding enzyme, a domain for unspecified protein binding, an enzyme with O-glucosyl hydrolase activity involved in carbohydrate metabolic process, membrane bound ion transporter with cation channel activity & ionotropic glutamate receptor activity, a transmembrane transporter with voltage-gated calcium channel activity, and calcium ion binding transmembrane ion transporter. A total of 3,943 core clusters contained 10 or more contigs, meaning that 71.32 % of the total core clusters consisted of less than 10 contigs. The majority of clusters fell within metabolic processes, cellular processes and catalytic activity.

4.1.2 Softcore transcriptome

A soft-core with four out of the five *Gambierdiscus* species examined was identified. The soft-core consisted of an additional 16,980 clusters (Table 2) of which 48 % were annotated (Suppl. table **??** & **??**). The most prolific cluster in the soft-core contained 163 contigs with unknown function, where *G. carpenteri G. lapillus*, *G. polynesiensis* and *G. holmesii* contained 50, 42, 41 & 30 contigs respectively. A further 5 clusters contained more than 100 contigs, four of which had GO annotations. Of the six clusters with over 100 contigs, none had representative contigs from *G. australes*. *G. australes* was absent from 86 % of the softcore clusters. In descending order of contigs, they matched to: a protein involved in selective protein binding, a protein involved in actin binding, a protein involved in calcium binding, and a protein with cysteine-type peptidase activity. In the soft-core, 14,035 clusters contained 10 or more contigs.

4.1.3 Accessory transcriptome

Clusters with single species representatives, or the accessory transcriptome to the five *Gambierdiscus* species examined, numbered 231,310 clusters. Of the unique clusters, only 15.23 % of clusters were annotated. Single species clusters from *G. australes*, *G. carpenteri* *G. lapillus*, *G. polynesiensis* and *G. holmesii* numbered 35,356, 62,494, 32,341, 60,796 & 41,350 clusters respectively (Table 2). The highest number of contigs in a unique cluster were 37, found in two clusters from *G. carpenteri*. One of these was annotated for RNA and metal ion binding activity. Of the unique clusters, 83.1 % contained only one contig and 97.8 % of clusters have 5 contigs or less.

4.1.4 Comparison of gene ontology annotations.

The GOs were split up into the three functional groups defined by the consortium: 1) Molecular processes (Figs. 4, 7, 10 & 13) defined as biochemical or a macromolecule directly interacting with other molecules, 2) Cellular components (Figs. 2, 6, 9 & 12) defined by the location within the cell where a molecular process takes place, and 3) Biological process (Figs 3, 5, 8 & 11) which is defined as a molecular machinery participating in the execution of the cell's genetic programming (e.g. cell division). GO basic is structured in a hierarchical manner, with parent and child terms where parent terms provide a broad overview of functionality and child terms are more specific than parent terms. For a general overview of functions present in each transcriptome, level 1 GO terms were elucidated (Figs. 3, 2 4, 8, 9 & 10). A more in depth query of the functions present in each transcriptome was conducted with a GO search of the child terms at level 2 (Figs. 5, 6, 7, 11, 12 & 13).

Level 1 GO annotations between *Gambierdiscus* species. The GO annotations found at level 1 between the species of *Gambierdiscus* were similar, with the exception of *G. australes* in several instances. GOs assigned part of catalytic activity in molecular processes (Fig. 4) as well as both the metabolic and cellular aspects in the biological processes (Fig. 3), *G. australes* was underrepresented as compared to the other four species. Within the molecular processes (Fig. 4), the most common annotation across all five species was for catalytic activity, followed by binding then transporter activities. Molecular carrier activity was only registered for *G. australes* and *G. carpenteri* with

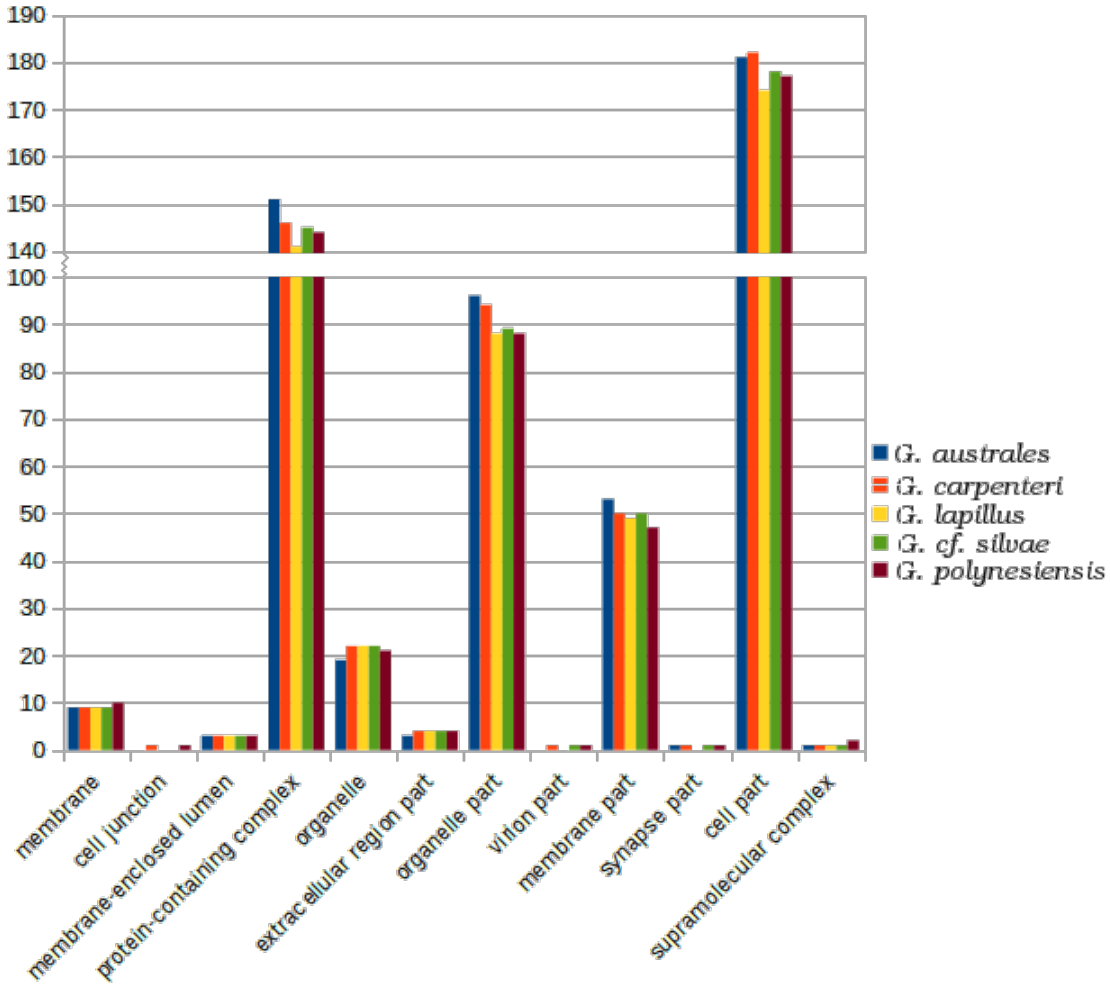

Figure 2: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 1.

1 annotation each. For GO annotations within the cellular processes (Fig. 2), the most common match was to cell parts followed by protein containing complexes then organelle parts. *G. carpenteri* and *G. polynesiensis* had one annotation each for cell junction activity. The highest number of GOs within biological processes matched to cellular processes (Fig. 3), closely followed by metabolic processes then biological regulation and localization. The lease represented biological GO annotation was related to growth with only one annotation for *G. holmesii* and *G. polynesiensis*.

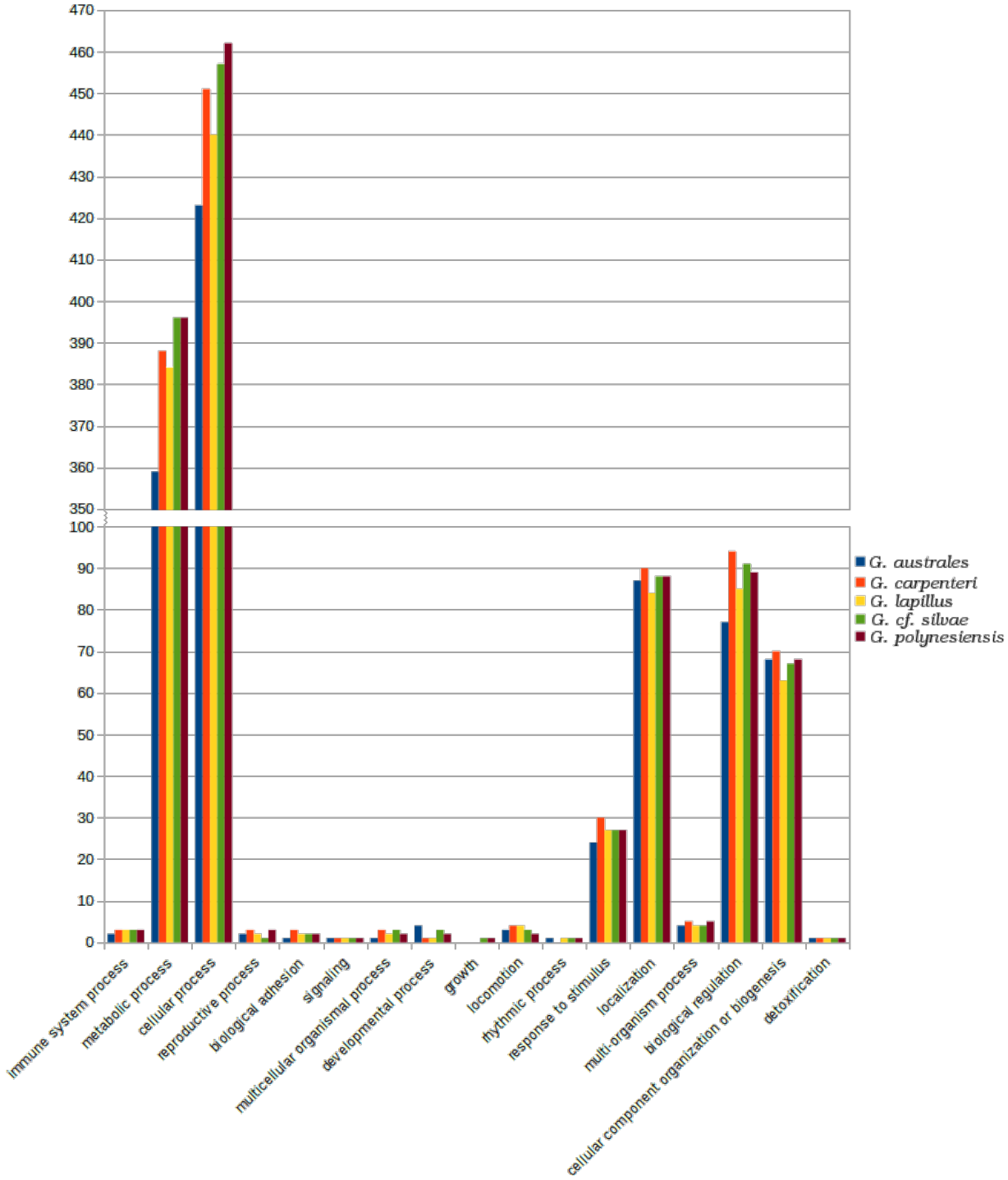

Figure 3: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 1.

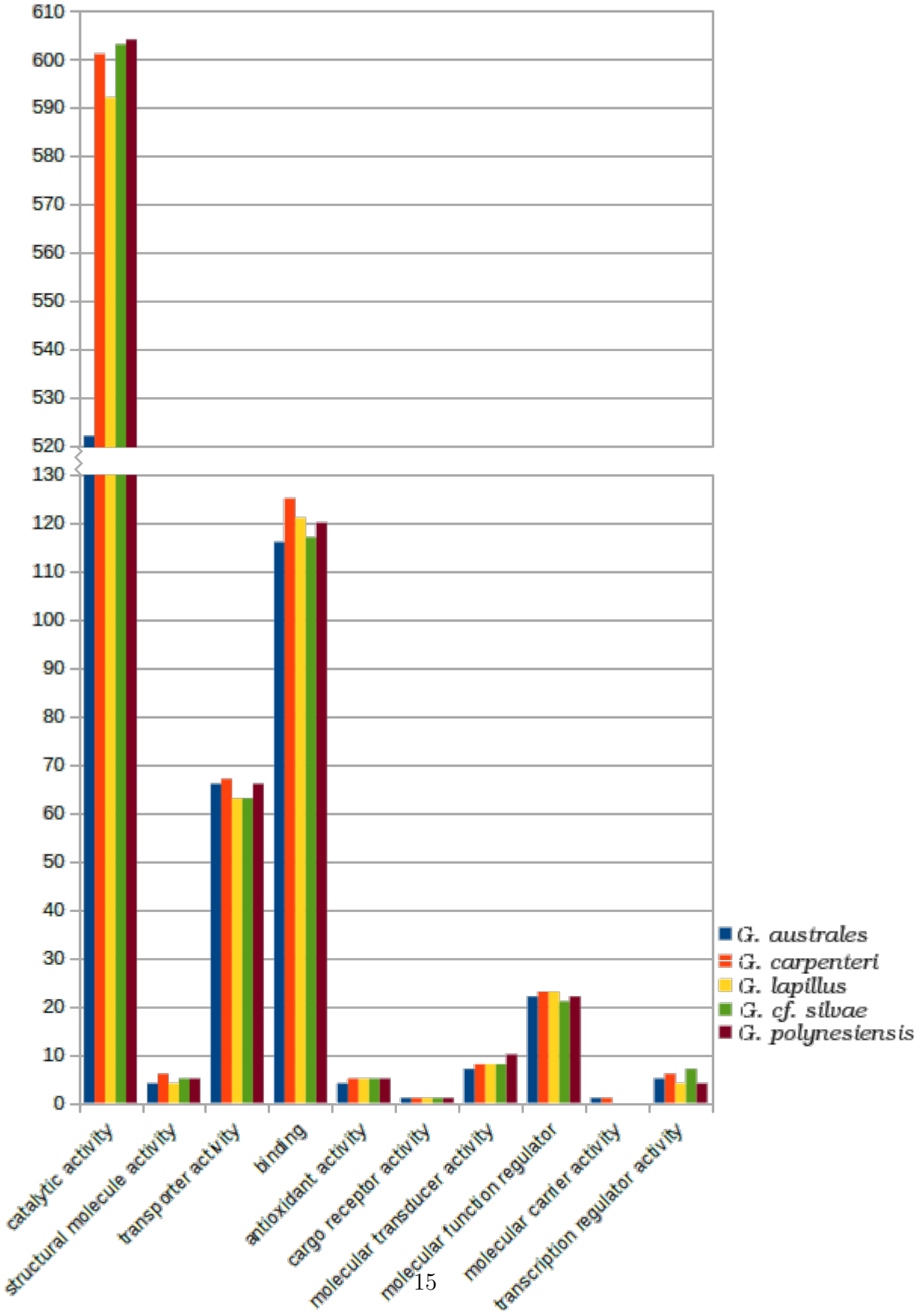

Figure 4: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 1.

Level 2 GO annotations between *Gambierdiscus* species. At level 2 of GO annotations, the difference between species becomes more apparent. While inter-species variations across the molecular, cellular and biological processes (Figs. 5, 6 & 7) are apparent, consistently *G. australes* is underrepresented or absent across all three processes. However, *G. australes* was the only species with a small number of GO annotations to nucleocytoplasmic carrier activity as well as general transcription initiation factor activity within the molecular processes. Again in the biological processes, *G. australes* was the only species matching to anatomical structure morphogenisis as well as movement within environment as part of symbiotic interaction. *G. holmesii* had a much higher representation of GO terms matching sperm-egg recognition. The most common molecular processes (Fig. 7) mapped to transferase activities, followed by hydrolase activity and oxidoreductase activity. For cellular processes (Fig. 6) the highest number of GOs was matched to intracellular parts, then intracellular organelle parts and membrane protein complexes. Organic substance metabolic processes, cellular metabolic processes and primary metabolic processes had the most GO annotation matches, for the biological processes group (Fig. 5).

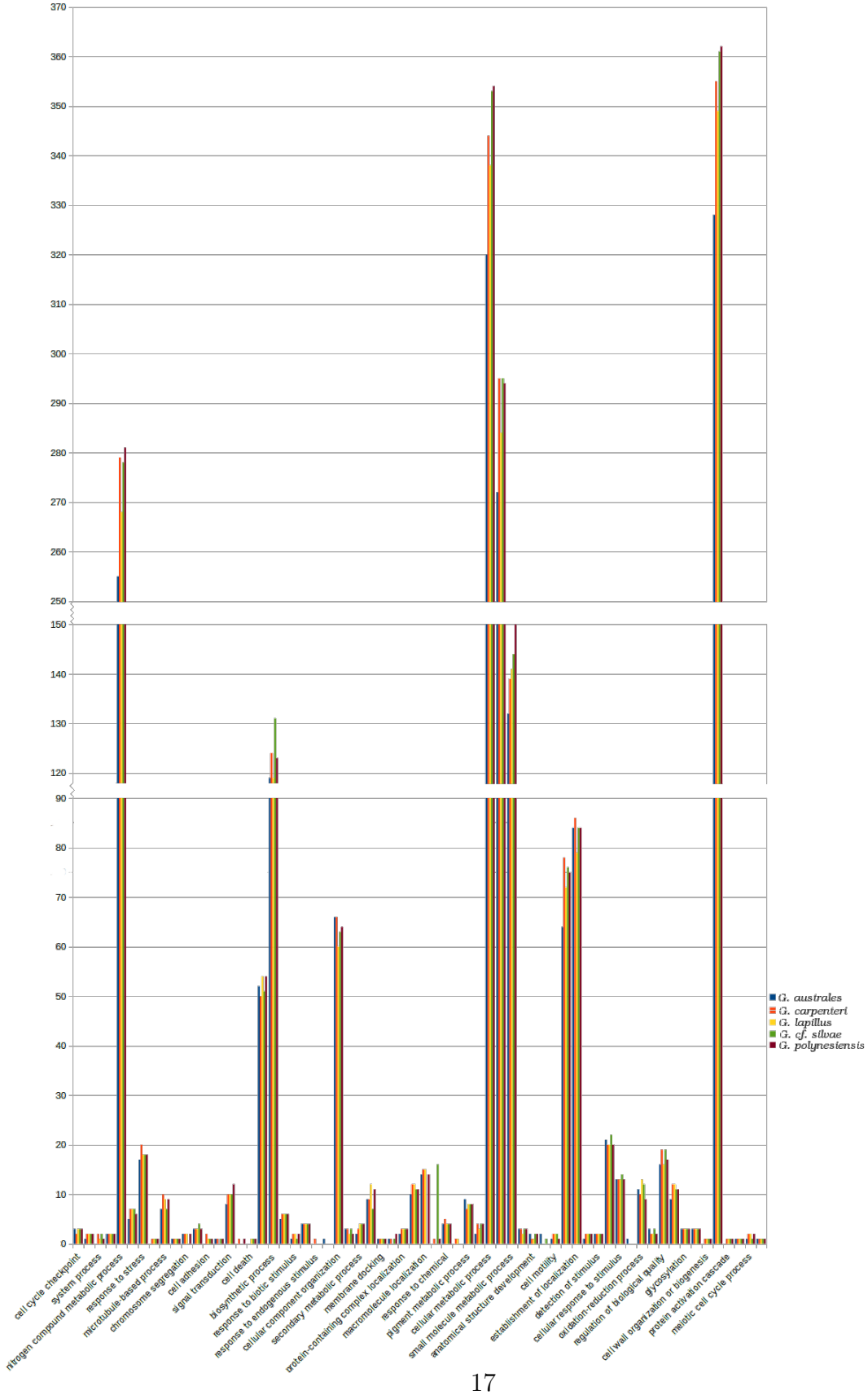

Figure 5: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 2.

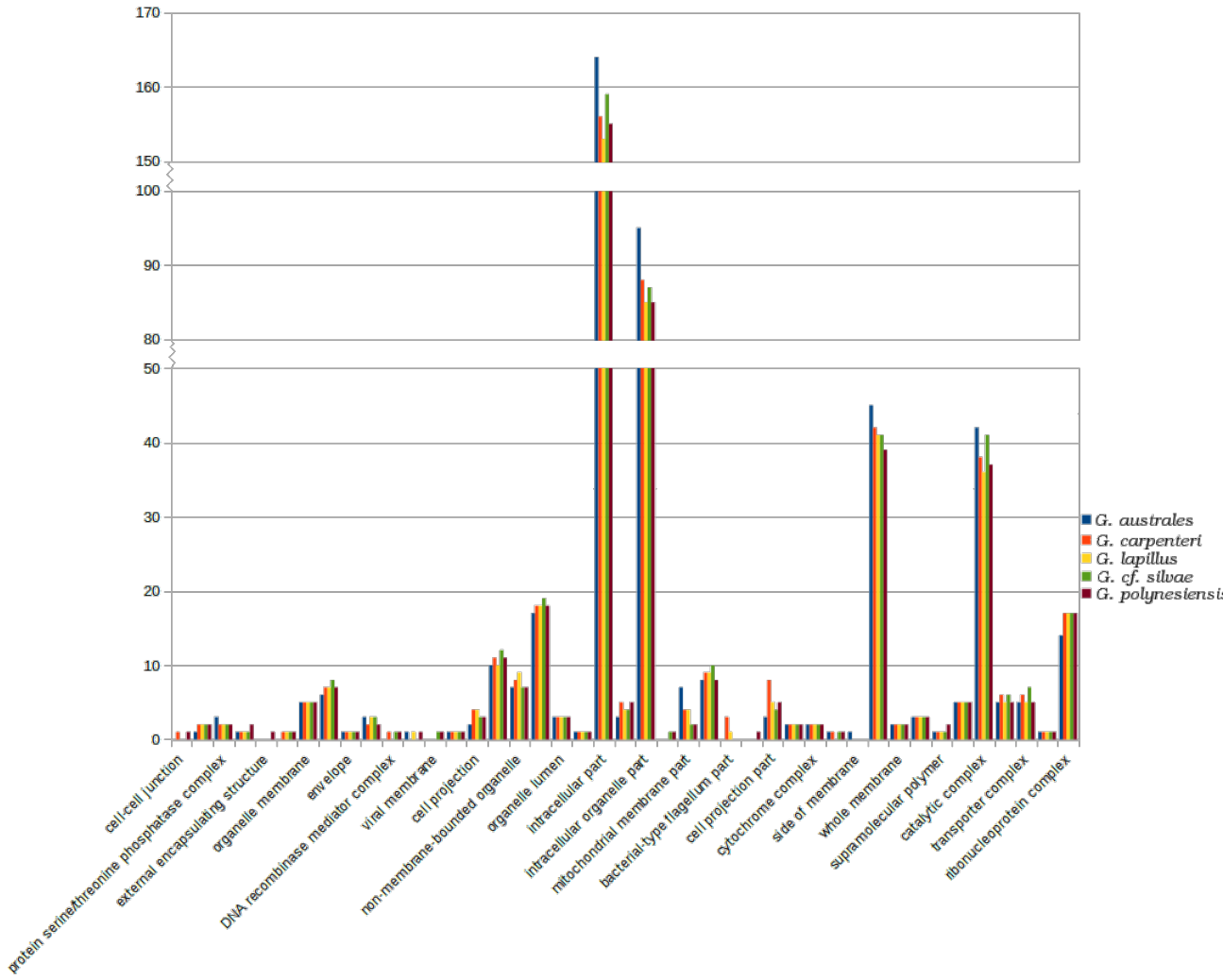

Figure 6: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 2.

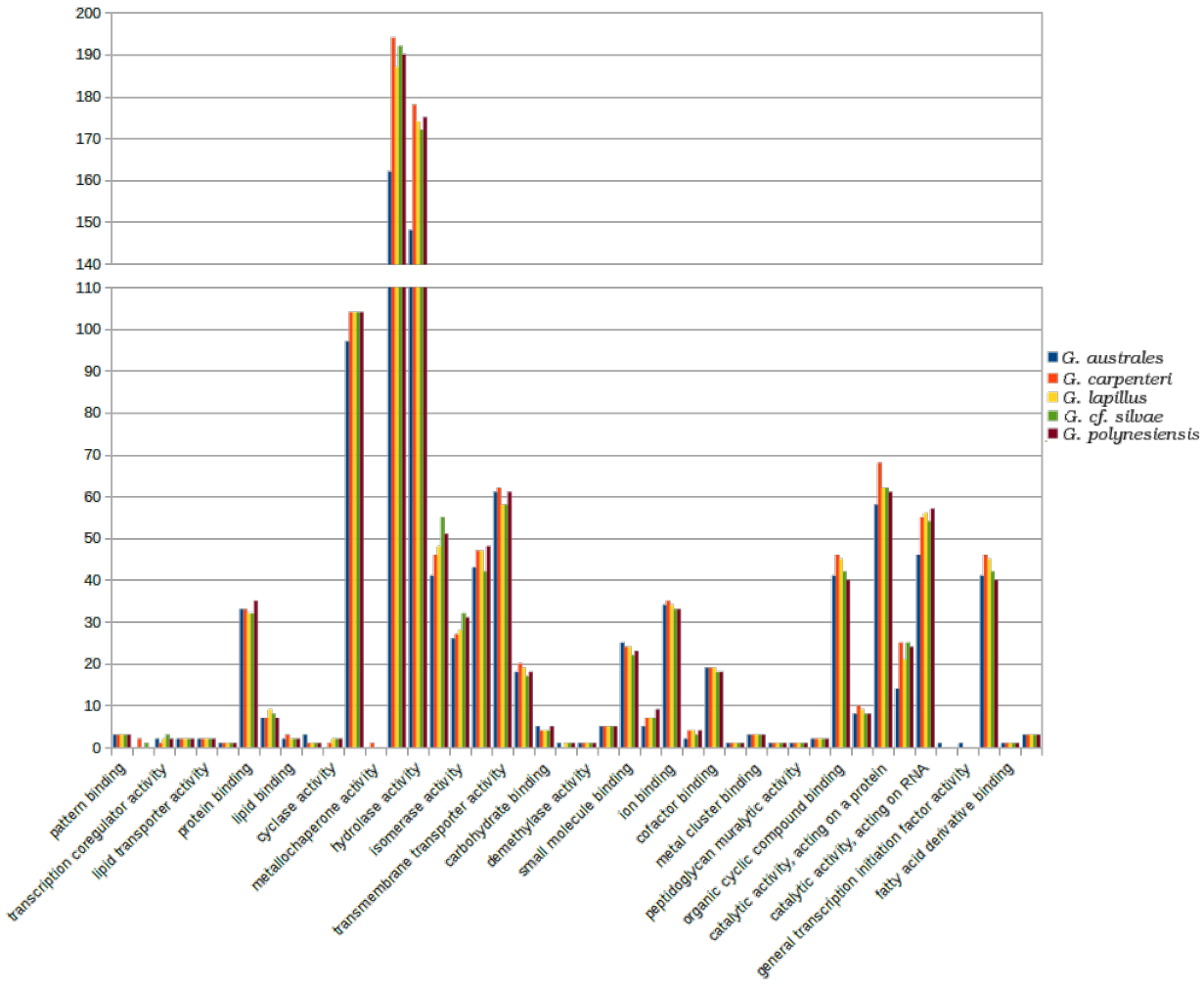

Figure 7: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 2.

Level 1 GO annotations for pan-transcriptomes. Similarity between the core, softcore and accessory clusters was consistent across the biological, cellular and molecular processes groups. Predominantly the unique clusters had a higher representation in each process with the exception for annotations matching to extracellular region parts and synpase parts within cellular processes (Fig. 9) as well as developmental processes within the biological processes (Fig. 8). GO annotations most commonly matched to catalytic activity then binding and transporter activities in the molecular processes (Fig. 10). Within the cellular processes, annotations predominantly matched to cellular parts, followed by protein-containing complexes and organelle parts (Fig. 8). For biological processes, the prevalent GO annotations matched to cellular processes, metabolic processes and localization (Fig. 8).

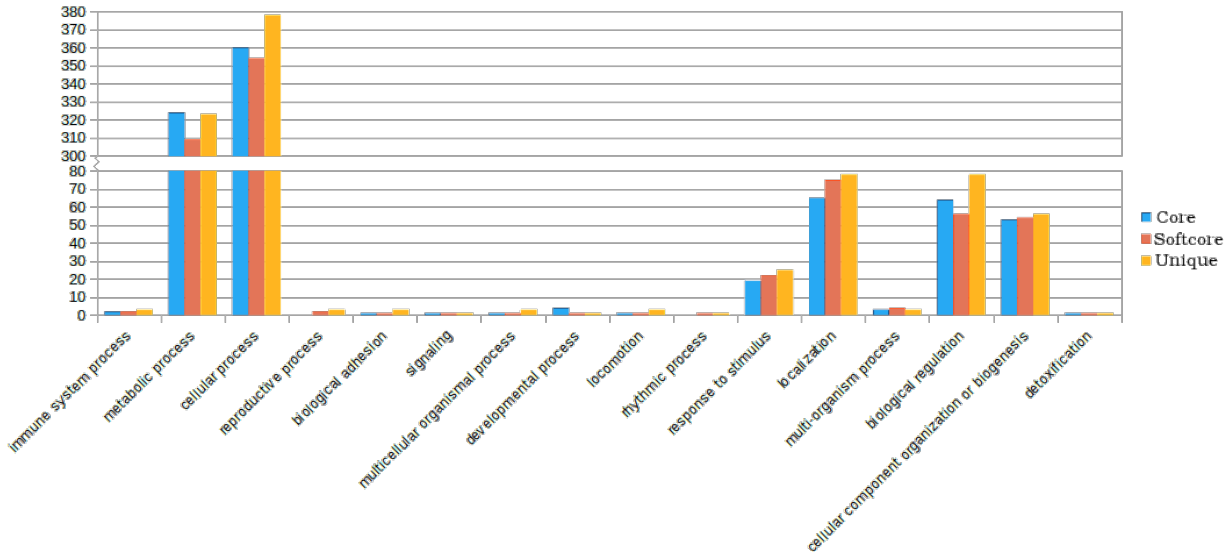

Figure 8: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 1.

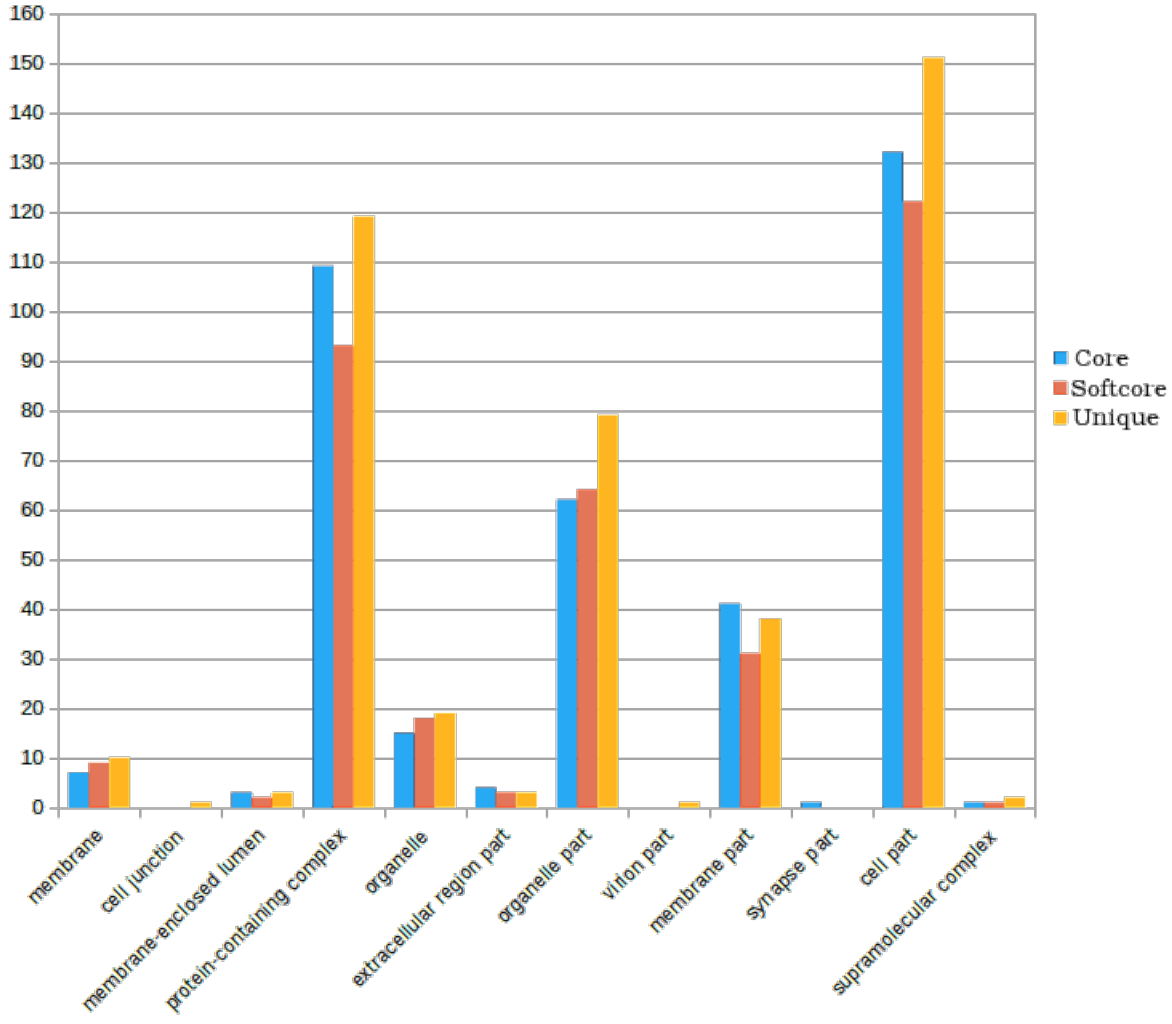

Figure 9: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 1.

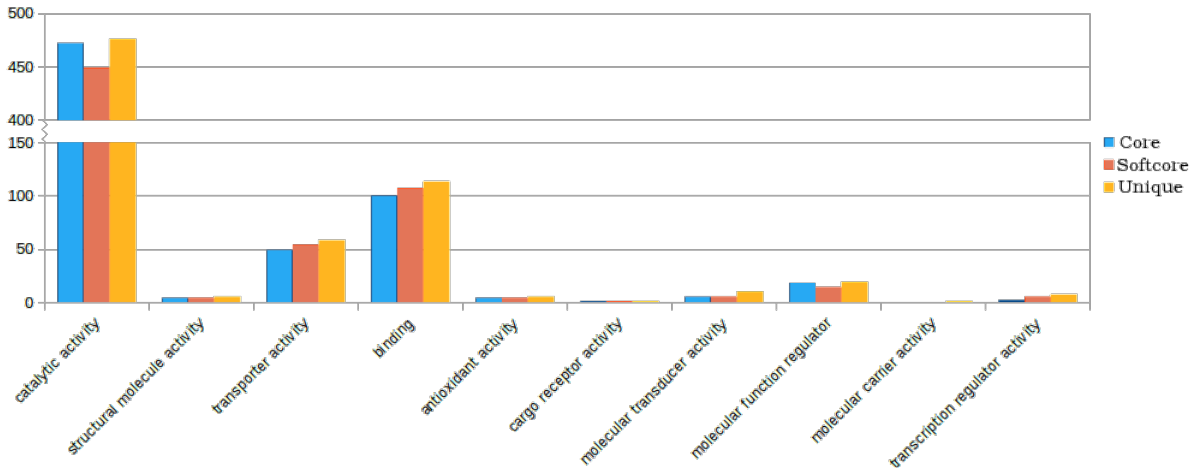

Figure 10: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 1.

Level 2 GO annotations for pan-transcriptomes. While differences between the biological, cellular and molecular processes were more distinctive at level 2, with the most common pattern among all three groupings of the core and unique clusters closely matching the number of GO terms with one or the other dominant, and the softcore clusters as less prevalent (Figs. 11, 12 & 13). Only the unique clusters had annotations matching to DNA binding transcription factor activity, metallochaperone activity and water binding in the molecular processes while the most common GO annotations for the accessory clusters matched to transferase, hydrolase and oxidoreductase activities in descending order (Fig. 13). Within the cellular processes, most annotations matched to intracellular parts, followed by intracellular organelle parts then membrane protein complexes (Fig. 12). Annotations solely from unique clusters were from cell-cell junction complexes, a viral membrane, contractile fibre parts, bacterial-type flagellum and an external encapsulating structure part. The biological processes annotations most commonly matched to organic substance metabolic processes, cellular metabolic processes and primary metabolic processes, in descending order (Fig. 11). Unique clusters were the only representatives for system processes, immune response, cell adhesion, cell death, sperm-egg recognition, cell motility and a protein activation cascade.

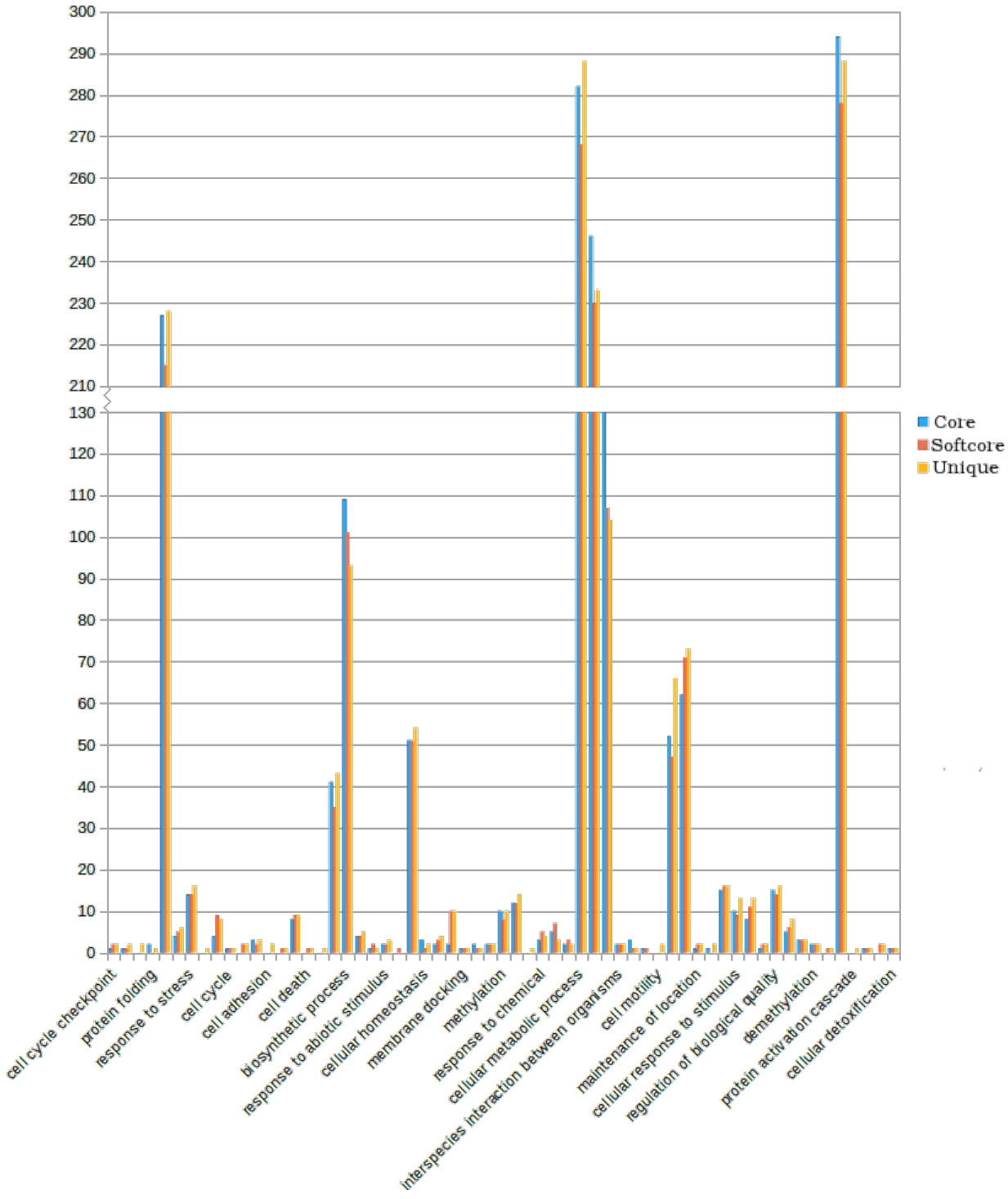

Figure 11: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 2.

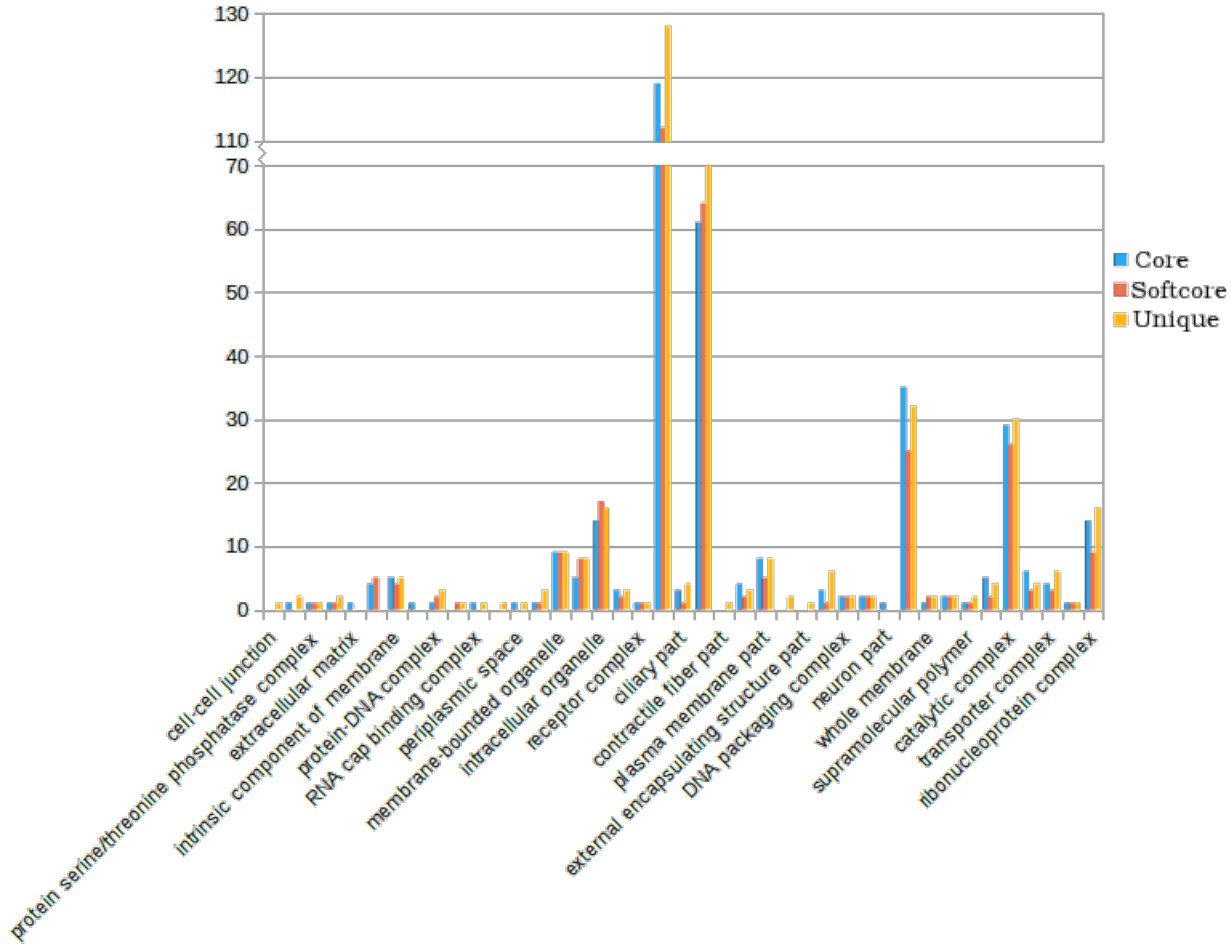

Figure 12: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 2.

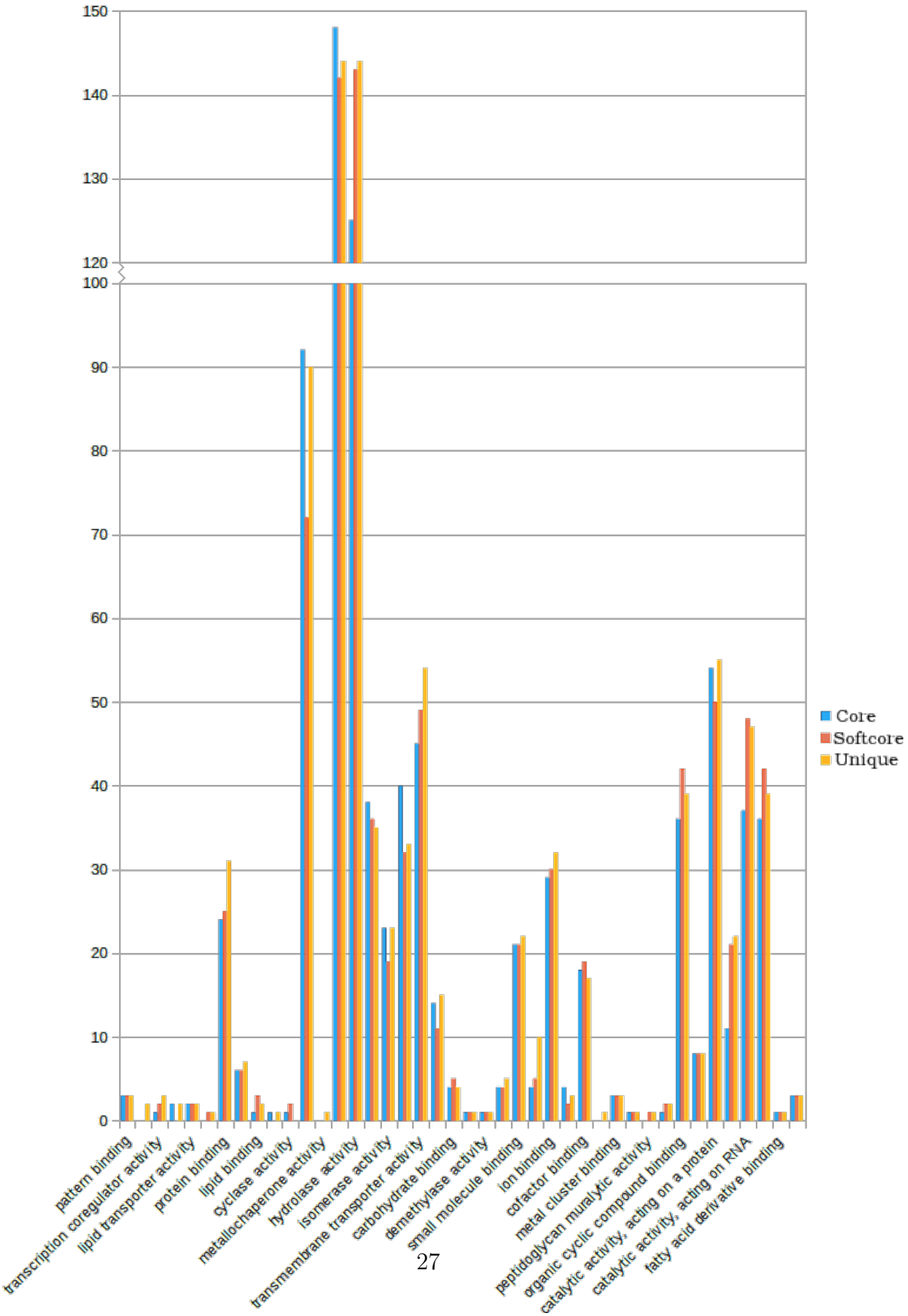

Figure 13: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 2.

4.2 Keto synthase active domain search

A total of 850 contigs were identified with KS domains which grouped into 314 clusters (Fig. 14). Nine clusters contained more than 10 contigs, with the highest number of 130 contigs from all species. Nine clusters contained 10 contigs or more, of which only two did not contain all the taxa examined. Fifty-seven of the 314 clusters contained contigs from multiple species, so 81.8 % of KS clusters were species specific while 78.7 % contained only a single contig (Fig. 14). The non-ciguatoxic *G. carpenteri* was absent from 73.6 % of the clusters. Of the clusters without *G. carpenteri*, none contained all four other species. However, one cluster contained *G. lapillus*, *G. polynesiensis* and *G. holmesii* with equally represented transcript numbers. Four contigs contained *G. polynesiensis* and *G. holmesii* only, one of which had a higher contig representation of *G. polynesiensis* than *G. holmesii*. *G. polynesiensis* was the only representative species in 71 clusters, of which three clusters contained 2 contigs and one cluster contained 3 contigs. *G. holmesii* was representative as the only species in 23 clusters, one of which contained 3 contigs while the other clusters contained single contigs. *G. australes*, *G. carpenteri* and *G. lapillus* were the solo representatives of 81, 39 & 35 KS clusters respectively.

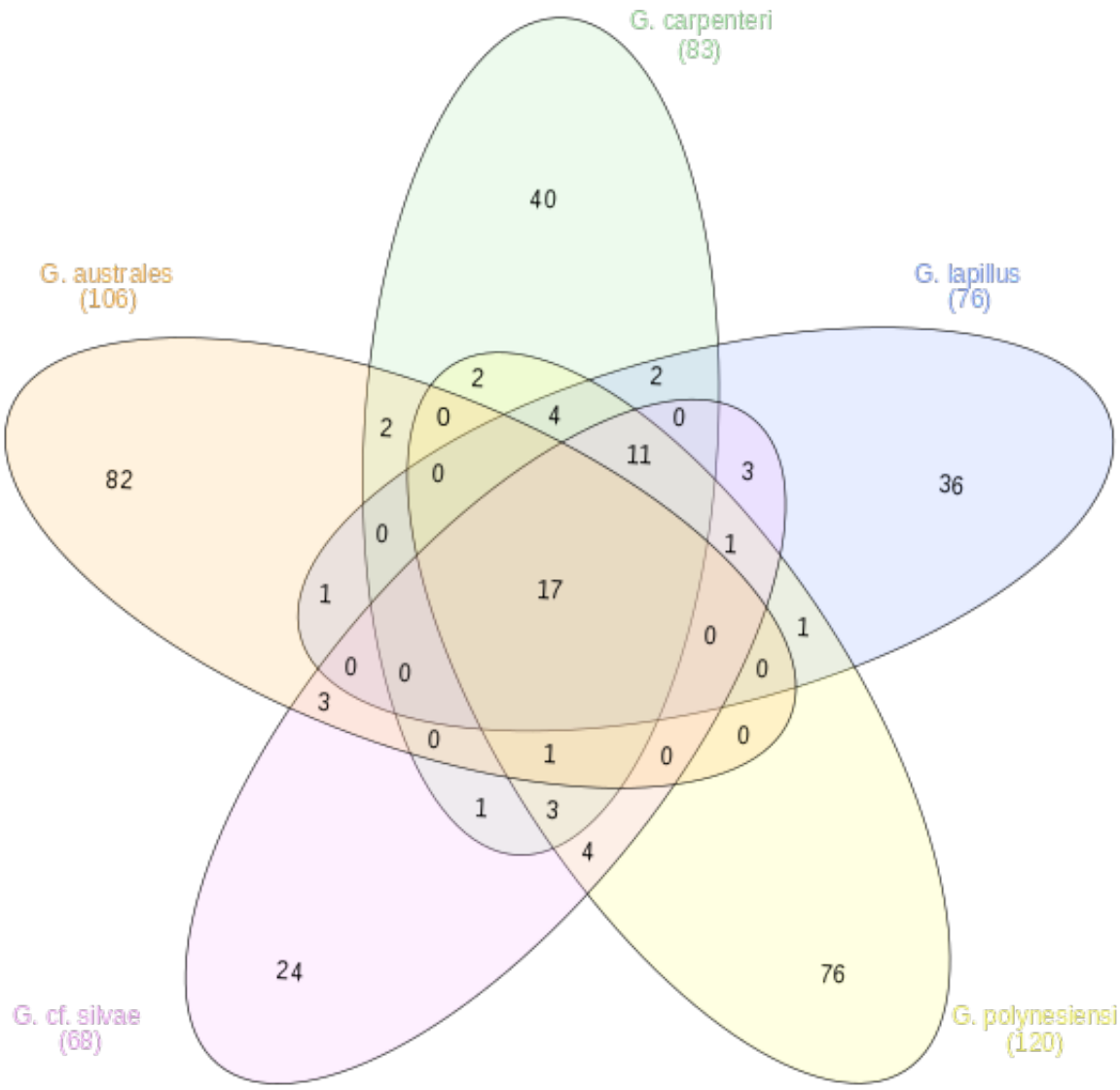

Figure 14: Venn diagram of ketosynthase containing clusters distributed across the five *Gambierdiscus* species. Core, accessory and inter-species ketosynthase cluster cross over depicted.

5 Discussion

Comparing five *Gambierdiscus* species revealed a core, soft-core and unique fraction in the transcriptomes. Further, differences between species with different toxin production characteristics were observed. The number of predicted peptides found in this study is high compared to a pan-transcriptome study of four prymnesiophyte algae [19]. The predicted peptides in the Koid et al. (2014) study ranged from around 25,000 to 56,000 peptides, where this study found a range from about 63,000 to 270,000 predicted peptides (Table 2). The lowest number of peptides was predicted in *G. australes*, similar to the findings in the prymnesiophyte algae study and both originating from the MMETSP, where sequencing occurred after a polyA-selection step for isolating eukaryotes [16, 19]. *G. holmesii* and *G. lapillus* predicted peptides numbered 132,688 and 111,862 respectively, while the highest number of predicted peptides were found in *G. carpenteri* and *G. polynesiensis* at 180,568 and 176,290 respectively. An exploration of the RNA sequencing libraries of two species of *Alexandrium* could not identify common polyadenylation signal sequence motifs [13]. The implication is that the polyadenylation signal is removed during mRNA processing, in at least some dinoflagellates [2, 13]. As the MMETSP transcriptomes had been sequenced after polyadenylation selection of transcripts, the comparatively low numbers of predicted peptides for *G. australes* as well as the prymnesiophyte algae could be accounted for by the loss of eukaryotic mRNA that had been processed to remove the polyadenylation signal.

The abundance of dinoSL was quite low (Table 2) compared to the abundance observed by Zhang et al. (2009) in *Amphidinium carterae*. Similar to this study, a low abundance of spliced leaders has been observed in other dinoflagellates [2, 8]. The function of the spliced leader, and hence whether this variation in observation between high and low abundance is species specific or due to assembly method employed, is as yet unknown. Interestingly, *G. holmesii* and *G. polynesiensis* had the most contigs with dinoSLs. These were sequenced with different read lengths and collected from different geographic locations, but are phylogenetically in the same *Gambierdiscus* sub-clade.

5.1 Core, softcore and accessory genes

The number of contigs and predicted peptides from this study was markedly less for *G. australes*, from the MMETSP dataset, in comparison to the transcriptome assemblies generated in **Chapter 4**. To accommodate for the low number of contigs from *G. australes*, the softcore spanned 4 of the 5 taxa. Noticeably, *G. australes* was absent from 86 % of the softcore dataset which indicates that a large proportion of the softcore is likely part of the *Gambierdiscus* pan-transcriptome core which was not captured in the *G. australes* sequencing. This is an example highlighting the value of a reference pan-transcriptome.

There was no distinct difference between the species GO annotations, with the exception of *G. australes*. This is as expected, as the species operate under similar nutritional modes and within similar temperature ranges [**? ?**] and were isolated from tropical or sub-tropical conditions, apart from *G. carpenteri* which was isolated from the temperate Merimbula region [22].

No difference was observed in the biological, cellular and molecular GO annotation groups at levels 1 and 2 between the core, softcore and accessory clusters. This is somewhat less expected as as it would be reasonable to predict a functional difference for genes unique to each species. Possible reasons could be that this observation only captures predicted peptides with GO annotations which were only 15.23 % of the 231,310 accessory clusters. This indicates that no functional match for over 196,000 of the accessory clusters could be found and no indication what their function might be could be extrapolated. As dinoflagellate sequencing and genome exploration is still comparably sparse, it follows that a lack of annotated references in the sequencing databases translates to low annotation rates. The need for annotations and a comparable reference transcriptome becomes apparent when investigating genetic elements as uncharacterized as these.

5.2 Ketosynthase domain detection

Between all five species, 850 contigs with KS domains were identified. These resolved into 314 clusters, of which 17 were shared among all species. The majority (81.8 %) resolved as unique clusters per species, ranging from 24 clusters for *G. holmesii* to 82 clusters for *G. australes*. The toxin profile for *G. polynesiensis* and the toxicity assay conducted

by Larsson et al. (2018) for *G. holmesii* indicate that these two species are the most ciguatoxic of the taxa included in this study. While *G. lapillus* displayed lower levels of ciguatoxicity in bioassays, the strain included here (HG4) was the least toxic of the *G. lapillus* strains tested in **Chapter 2**. Four KS clusters included *G. holmesii* and *G. polynesiensis* only, one of which contained 4 contigs from *G. polynesiensis* and 1 from *G. holmesii*. A further 22 and 75 clusters only contained *G. holmesii* and *G. polynesiensis* contigs respectively. These clusters could be of interest for further investigation into KS domains involved in CTX synthesis.

5.3 Areas for possible improvement in this study

This chapter presents a novel approach to analyzing *Gambierdiscus* transcriptomes and possible avenues for investigation for ciguatoxin production pathways. However, there are several aspects that can be improved upon with future studies.

Contaminants in dataset. The RNA-seq libraries were constructed from whole RNA seq runs with non-axenic cultures, hence it is likely that bacterial RNA is a subset of the analysis. It is unlikely that the same contamination persists in the core and softcore clusters, i.e. across all four to five species collected from Australia, the Cook Islands and over a 10 year time span. However the unique clusters could well be contaminated with non-eukaryotic contigs, which is indicated by some of the unique clusters only annotations mapping to a bacterial flagellum and a viral capsid. Hence it would be pertinent to devise a method to separate bacterial from eukaryotic contigs post sequencing.

Coverage of taxa. As with bacterial pan-transcriptomics, eukaryotic studies reveal a correlation between the number of species and strains included for refining the core transcriptome and expanding the unique fraction of the transcriptome [14, 19, 36]. For *Gambierdiscus*, including several strains of a species is essential for the discovery of a non-ciguatoxic *G. polynesiensis* strain [33]. Variability in morphology and toxicity has also been observed in *G. lapillus* [20]. While this study sought to cover taxa from the three main *Gambierdiscus* clades, an increase in species and strain coverage is highly likely to impact the resolution of the core and unique portions especially as the sequencing coverage of the *G. australes* transcriptome is less in depth than the other four species

(Table 2). Hence due to limited species and the singular strain per species coverage, this represents an exploratory study for establishing a pan-transcriptome for *Gambierdiscus* which should be improved upon.

PKS active domain search. PKS complexes consist of a number of active domains that synthesize and manipulate the polyketide backbone. The KS domain is but one of several essential domains for a functional polyketide. The search for active domains should be extended to include other domains for comparison in the search for the ciguatoxin production pathways.

Conclusion

Protists represent a large section of the tree of life and are involved in vital geochemical cycling, symbiotic and toxic relationships in their environment. Yet due to the convoluted nature of their genomes, querying the genetic content of these organisms is fraught with obstacles. This study presents a pan-transcriptome for the genus *Gambierdiscus*, some species of which are involved in CTX production. The approach opens up an alternative avenue for investigation of the differences and similarities of toxic *Gambierdiscus* species in general, and specifically in regard to the toxin production pathway(s). This study provides a starting point for *Gambierdiscus* pan-transcriptomic exploration and a rudimentary reference for future sequencing efforts. It is recommended that this dataset is expanded to encompass both more *Gambierdiscus* species and strains to crystallize the genus' pan-transcriptome in more detail.

6 References

[1] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene Ontology: tool for the unification of biology. *Nature genetics 25*, 1 (2000), 25.

[2] Bachvaroff, T. R., and Place, A. R. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. *PLoS One 3*, 8 (2008), e2929.

[3] Cerveau, N., and Jackson, D. J. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC bioinformatics 17*, 1 (2016), 525.

[4] Chinain, M., Darius, H. T., Ung, A., Cruchet, P., Wang, Z., Ponton, D., Laurent, D., and Pauillac, S. Growth and toxin production in the ciguatera-causing dinoflagellate *Gambierdiscus polynesiensis* (Dinophyceae) in culture. *Toxicon 56*, 5 (2010), 739–750.

[5] Consortium, G. O. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research 45*, D1 (2016), D331–D338.

[6] Eddy, S., and Wheeler, T. HMMER: biosequence analysis using profile hidden Markov models, 2015. hmmer.org/.

[7] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics 28*, 23 (2012), 3150–3152.

[8] Guo, R., and Ki, J.-S. Spliced leader sequences detected in EST data of the dinoflagellates *Cochlodinium polykrikoides* and *Prorocentrum minimum*. *Algae 26*, 3 (2011), 229–235.

[9] Haas, B., and Papanicolaou, A. TransDecoder (find coding regions within transcripts), 2016.

[10] Harke, M. J., Juhl, A. R., Haley, S. T., Alexander, H., and Dyhrman, S. T. Conserved transcriptional responses to nutrient stress in bloom-forming algae. *Frontiers in microbiology 8* (2017), 1279.

[11] He, F., and Maslov, S. Pan-and core-network analysis of co-expression genes in a model plant. *Scientific reports 6* (2016), 38956.

[12] Heberle, H., Meirelles, G., da Silva, F., Telles, G., and Minghim, R. Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics 16* (2015), 169.

[13] Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A., and John, U. Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. *PLoS One 6*, 12 (2011), e28012.

[14] Jin, M., Liu, H., He, C., Fu, J., Xiao, Y., Wang, Y., Xie, W., Wang, G., and Yan, J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific reports 6* (2016), 18936.

[15] Kahlke, T. GOSUM: Gene Ontology Summarizer version 0.1, 2018. http://doi.org/10.5281/zenodo.1344306.

[16] Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PloS one* (2014).

[17] Kohli, G. S., Campbell, K., John, U., Smith, K. F., Fraga, S., Rhodes, L. L., and Murray, S. A. Role of modular polyketide synthases in the production of polyether ladder compounds in ciguatoxin-producing *Gambierdiscus polynesiensis* and *G. excentricus* (Dinophyceae). *Journal of Eukaryotic Microbiology* (2017).

[18] Kohli, G. S., John, U., Figueroa, R. I., Rhodes, L. L., Harwood, D. T., Groth, M., Bolch, C. J., and Murray, S. A. Polyketide synthesis genes

associated with toxin production in two species of gambierdiscus (dinophyceae). *BMC genomics 16*, 1 (2015), 410.

[19] Koid, A. E., Liu, Z., Terrado, R., Jones, A. C., Caron, D. A., and Heidelberg, K. B. Comparative transcriptome analysis of four prymnesiophyte algae. *PLoS One 9*, 6 (2014), e97801.

[20] Kretzschmar, A. L., Verma, A., Harwood, T., Hoppenrath, M., and Murray, S. Characterization of gambierdiscus lapillus sp. nov.(gonyaulacales, dinophyceae): A new toxic dinoflagellate from the great barrier reef (australia). *Journal of phycology 53*, 2 (2017), 283–297.

[21] Lapierre, P., and Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends in genetics 25*, 3 (2009), 107–110.

[22] Larsson, M. E., Laczka, O. F., Harwood, D. T., Lewis, R. J., Himaya, S., Murray, S. A., and Doblin, M. A. Toxicology of *Gambierdiscus* spp.(Dinophyceae) from tropical and temperate Australian waters. *Marine drugs 16*, 1 (2018), 7.

[23] Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology 32*, 10 (2014), 1045.

[24] Meyer, J. M., Rödelsperger, C., Eichholz, K., Tillmann, U., Cembella, A., McGaughran, A., and John, U. Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate *Azadinium spinosum*, with emphasis on polykeitde synthase genes. *BMC genomics 16*, 1 (2015), 27.

[25] Munday, R., Murray, S., Rhodes, L. L., Larsson, M. E., and Harwood, D. T. Ciguatoxins and maitotoxins in extracts of sixteen gambierdiscus isolates and one fukuyoa isolate from the south pacific and their toxicity to mice by intraperitoneal and oral administration. *Marine drugs 15*, 7 (2017), 208.

[26] Murray, S. A., Suggett, D. J., Doblin, M. A., Kohli, G. S., Seymour, J. R., Fabris, M., and Ralph, P. J. Unravelling the functional genetics of

dinoflagellates: a review of approaches and opportunities. *Perspect. Phycol 3*, 1 (2016), 37–52.

[27] Pawlowiez, R., Morey, J., Darius, H., Chinain, M., and Van Dolah, F. Transcriptome sequencing reveals single domain Type I-like polyketide synthases in the toxic dinoflagellate *Gambierdiscus polynesiensis*. *Harmful Algae 36* (2014), 29–37.

[28] Plissonneau, C., Hartmann, F. E., and Croll, D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC biology 16*, 1 (2018), 5.

[29] Posnien, N., Zeng, V., Schwager, E. E., Pechmann, M., Hilbrant, M., Keefe, J. D., Damen, W. G., Prpic, N.-M., McGregor, A. P., and Extavour, C. G. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. *PLoS One 9*, 8 (2014), e104885.

[30] Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. InterProScan: protein domains identifier. *Nucleic acids research 33*, suppl_2 (2005), W116–W120.

[31] Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature 499*, 7457 (2013), 209.

[32] Rhodes, L. L., Smith, K. F., Munday, R., Selwood, A. I., McNabb, P. S., Holland, P. T., and Bottein, M.-Y. Toxic dinoflagellates (Dinophyceae) from Rarotonga, Cook Islands. *Toxicon 56*, 5 (2010), 751–758.

[33] Rhodes, L. L., Smith, K. F., Murray, S., Harwood, D. T., Trnski, T., and Munday, R. The epiphytic genus *Gambierdiscus* (Dinophyceae) in the Kermadec Islands and Zealandia regions of the southwestern Pacific and the associated risk of ciguatera fish poisoning. *Marine drugs 15*, 7 (2017), 219.

[34] Ryan, D. E., Pepper, A. E., and Campbell, L. De novo assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*. *BMC genomics 15*, 1 (2014), 888.

[35] Song, G., Dickins, B. J., Demeter, J., Engel, S., Dunn, B., and Cherry, J. M. AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*. *PLoS One 10*, 3 (2015), e0120671.

[36] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences 102*, 39 (2005), 13950–13955.

[37] Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. Ten years of pan-genome analyses. *Current opinion in microbiology 23* (2015), 148–154.

[38] Vinuesa, P., and Contreras-Moreira, B. Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids. In *Bacterial Pangenomics*. Springer, 2015, pp. 203–232.

[39] Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., and Lin, S. Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences 104*, 11 (2007), 4618–4623.