

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

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

1 Abstract

Species of the genus *Gambierdiscus* produce Ciguatoxins (CTXs), the causative agent of ciguatera fish poisoning, a potentially debilitating seafood borne illness. Species of *Gambierdiscus* possess very large genomes, 32 - 35 Gbp, and, as with other dinoflagellates, possess unique genomic characteristics, such as highly repetitive and complex genome architecture. The exact toxins produced by species of *Gambierdiscus* remain largely unclear. It has been verified using LCMS on multiple strains that the species *Gambierdiscus polynesiensis* produces analogs of CTXs. Other species appear to produce maitotoxins, gambierol, and other uncharacterised toxins. An understanding of the evolution of *Gambierdiscus* and their toxins requires information regarding their genetics. Transcriptomic sequencing is a feasible alternative to genome sequencing. In this study, we generated de novo RNA-seq libraries for *Gambierdiscus polynesiensis*, *Gambierdiscus carpenteri*, *Gambierdiscus holmesii* and *Gambierdiscus lapillus*, compared these to a previously sequenced *Gambierdiscus australes*, to discover a set of core genes shared by all species. We present a *Gambierdiscus* core transcriptome, which might be used to investigate candidate genes related to toxin production.

To do:

- re-structure as per Tim's comments
- incl Sammy's comments

2 Introduction

The challenge of protist *de novo* sequencing projects lies in assessing the adequacy and completeness of sequencing as well as library processing and assembly methods employed, without a well annotated reference. This issue is particularly prevalent in dinoflagellates, whose expansive and complex genetics tend to be a barrier to genomic sequencing. As an alternative to wrangling with dinoflagellate genomes, transcriptomes are used as to explore their genetics. This is due to the apparent presence of uncharacterized genetic mechanism(s) which seem to leave protein synthesis regulation to the post-transcriptional stage, thus with the effect that mRNA gives an approximation of genomic content. An indication of these regulatory mechanisms comes from a number of direct previous observations. Harke et al. (2017) cultured *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. 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 [11]. 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 [11]. As these organisms are relatively difficult to culture and extract RNA, until the MMEPTSP the number of marine eukaryotic transcriptomes was sparse. When searching for *Gambierdiscus* on NCBI's SRA database 5 relevant projects were found in addition to the MMEPTSP 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 *Cryptothecodinium*, which are arguably not part of the gonyaulacales (see **chapter 4**). Searching for members of the phylum dinoflagellates calls a further 84 projects. Despite their ecological relevance for nutrient cycling, DMSP production, coral symbiosis and neurotoxin production (for a review see [30]), the paucity of sequencing data, even with the MMEPTSP dataset, is evident. This is further confounded to a large proportion of dinoflagellate transcriptomes

sharing no known similarity to other described proteins or domains compared to known databases. When compared to NCBI's nr database, the proportion of contigs with no known match was 60 % for *Azadinium spinosum* [27], over 50 % for *G. australes* & *G. belizeanus* [19], 57.9 % for *G. excentricus* [18], 63 % for *G. polynesiensis* [18, 31], and 55 - 57 % for *Karenia brevis* [38].

The concept of a reference genome, or transcriptome, allows for direct comparison of genome/transcriptome sequencing to a 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 [41, 42]. An alternative approach to a reference genome was proposed - that of a core-genome common to all strains, and a pan-genome which is transitory. An extrapolation of this study by Tettelin et al. (2005), which showed that 1.5 % of the genome was novel between 8 strains of *Streptococcus* predicted based on mathematical models that for every new strain sequenced 22 novel genes are predicted to be discovered [26]. Since then the core- and pan-genome, or transcriptome, concept has been adopted for eukaryotes also, with the realisation that the transient genomic content holds true when multiple strains of a species are sequenced (e.g. [14, 24, 32, 33, 35, 39]). Further to exploring the shared and transient genetic components within a genus, pan and core analyses have been conducted for higher taxonomic levels, commonly within genus though also at much higher levels, such as the gene frequency of *Eubacteria* within the super kingdom inter-species pan and core analysis have also been conducted [12, 14, 20, 22, 42].

Five transcriptomes of *Gambierdiscus* were compared in this study with the aim of providing a pan-transcriptomic baseline *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 been implicated in MTX production via bioassays, while *G. carpenteri* did not register for CTX-like activity in a bioassay [23]). The toxin profiles registered all species apart from *G. carpenteri* as an MTX producer, 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, unique portion of the transcriptome. The results in this study

could provide an avenue of investigation of quering 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	<i>G. australes</i>	<i>G. carpenteri</i>	<i>G. lapillus</i>	<i>G. polyne-siensis</i>	<i>G. holmesii</i>
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	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	[17, 29, 36]	[23]	[21, 23]	this study, [?]	[21, 23]

3 Methods

Scripts used for this project are available on Github under `hydrahamster/pan-tran`. Venn diagrams were created with InteractiVenn [13].

3.1 Transcriptome acquisition

Species of *Gambierdiscus* used in this chapter are summarized in Table 1. Toxicity and toxin profile reports are specific to the strains used as inter-species variation in toxin production was recently reported [23, 37], unless noted otherwise. The *G. polynesiensis* toxin profile was elucidated by Tim Harwood at the Cawthron institute with the same methodology as for *G. lapillus* in **Capter 2**. Seq libraries were assembled as per the transcriptome assembly subsection in the methods of **chapter 4**, without `diginorm`.

3.2 Spliced leader search

The spliced leader sequences reported by Zhang et al. (2007) were used to build a `hmmer` library [?]. 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, 8]. `Transdecoder` was used to predict coding regions on the clustered nucleotide sequences [10]. Protein clusters were annotated with `Interproscan v5.27` with local lookup server [34]. Protein clusters were processed to include the species of origin instead of the `TRINITY` tag and concatenated for input to `get_homologues` [43]. 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 resulting core-, softcore- and unique-clusters were matched with their `interpro` annotations and `GO` terms were queried with `GOSUM`

against the basic Gene Ontology (GO) database [1, 4, 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. The KS domains found were aligned with MUSCLE with a maximum of 8 iterations [7]. Maximum likelihood (ML) inference was run with the KS alignments using RaxML [40] with the -PROTGAMMAILGF flags on the University of Technology Sydneys High-performance computing cluster (HPCC)

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).

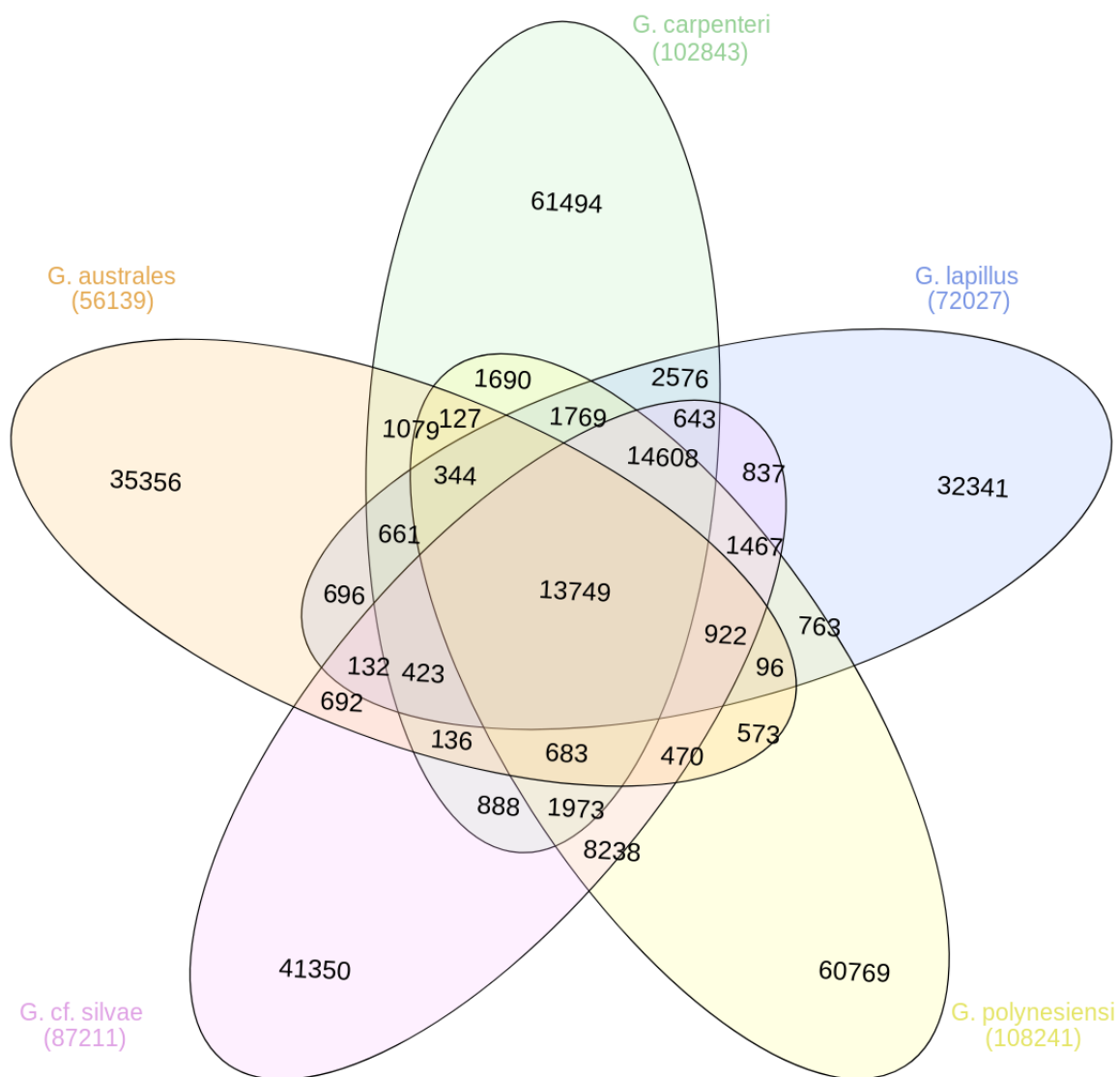


Figure 1: Venn diagram of species distribution across clusters.

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

Species	<i>G. aus- trales</i>	<i>G. carpenteri</i>	<i>G. lapillus</i>	<i>G. polyne- siensis</i>	<i>G. holme- sii</i>
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 (Transde- coder)	63,299	180,568	111,862	176,290	132,688
Contigs anno- tated (Inter- pro 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 clus- ters	35,356	61,494	32,341	60,769	41,350

Tim: It seems kinda conspicuous that the unique clusters of *G. carpenteri* & *G. poly* are almost twice the number of *G. lapillus* and *G. holmesii*, the first two were sequenced together with 150bp read length while the other two had 75bp read length during sequencing. Does this seem odd to you too?

4.1.1 Core transcriptome

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 5 & 6). The highest number of contigs in any core cluster was 180 cluster of unknown function with 23, 45, 32, 31 and 49 from *G. australes*, *G. carpenteri*, *G. lapillus*, *G. polynesiensis* and *G. holmesii* respectively. 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: 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, so 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 with %, % and % of annotated clusters respectively. **Tim** - so adding up the lvl1 gosum counts for bio, cell and molec doesn't add up to the total annotated clusters.. am I correct in thinking that this is because annotations can go to other functions too?

4.1.2 Softcore transcriptome

A softcore with 4 out of the five *Gambierdiscus* species examined was identified. The softcore consisted of an additional 16,980 clusters (Table 2) of which 48 % were annotated (Suppl. table 5 & 6). The most prolific cluster in the softcore 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 representatives 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. Of the softcore, 14,035 clusters contained 10 or more contigs.

4.1.3 Unique part of the transcriptome

Clusters with single species representatives, or the pan-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 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. For GOs assigned part of catalytic activity in molecular processes (Fig. 10) as well as both the metabolic and cellular processes in the biological processes (Fig. 8), *G. australes* was underrepresented. Within the molecular processes (Fig. 10), the most common annotation was for catalytic activity, followed by binding then transporter activities. Molecular carrier activity was only registered for *G. australes* and *G. carpenteri* with 1 annotation each. For GO annotations within the cellular processes (Fig. 9), the most common match was to cell parts followed by protein containing complexes then organelle parts. Only *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. 8), closely followed by metabolic processes then biological regulation and localization. The least represented biological GO annotation was related to growth with only one annotation for *G. holmesii* and *G. polynesiensis*.

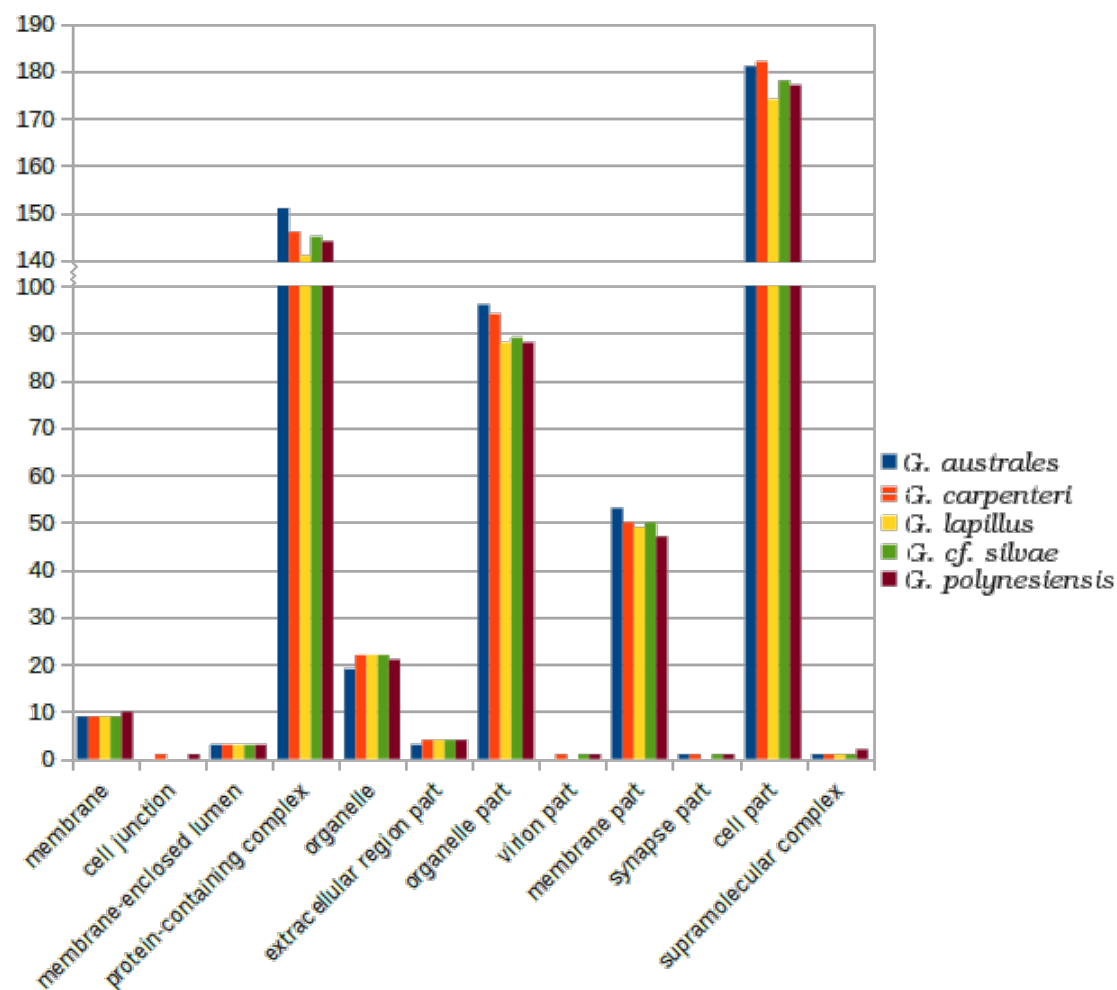


Figure 2: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 1 from Suppl. table 3.

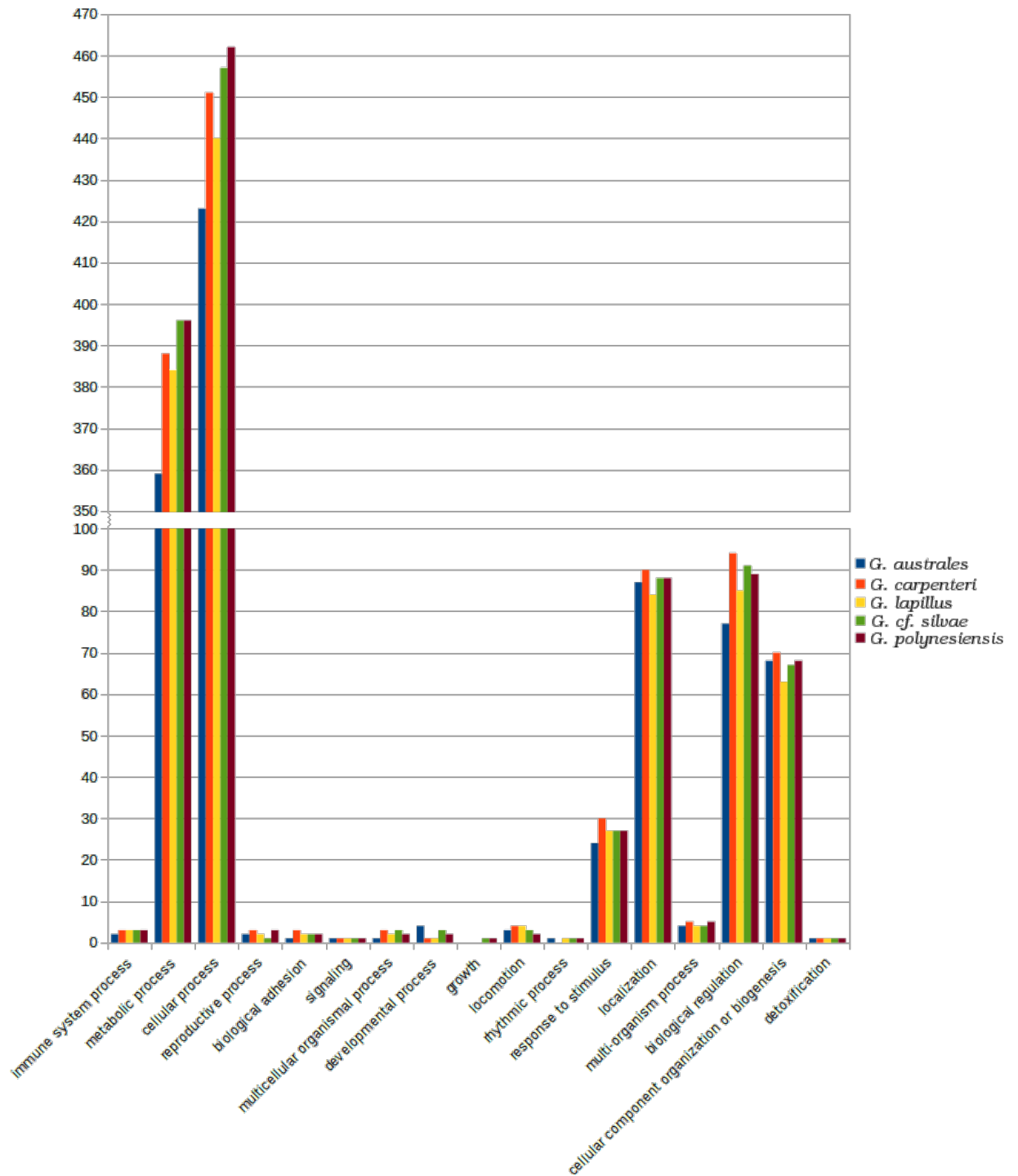


Figure 3: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 1 from Suppl. table 3.

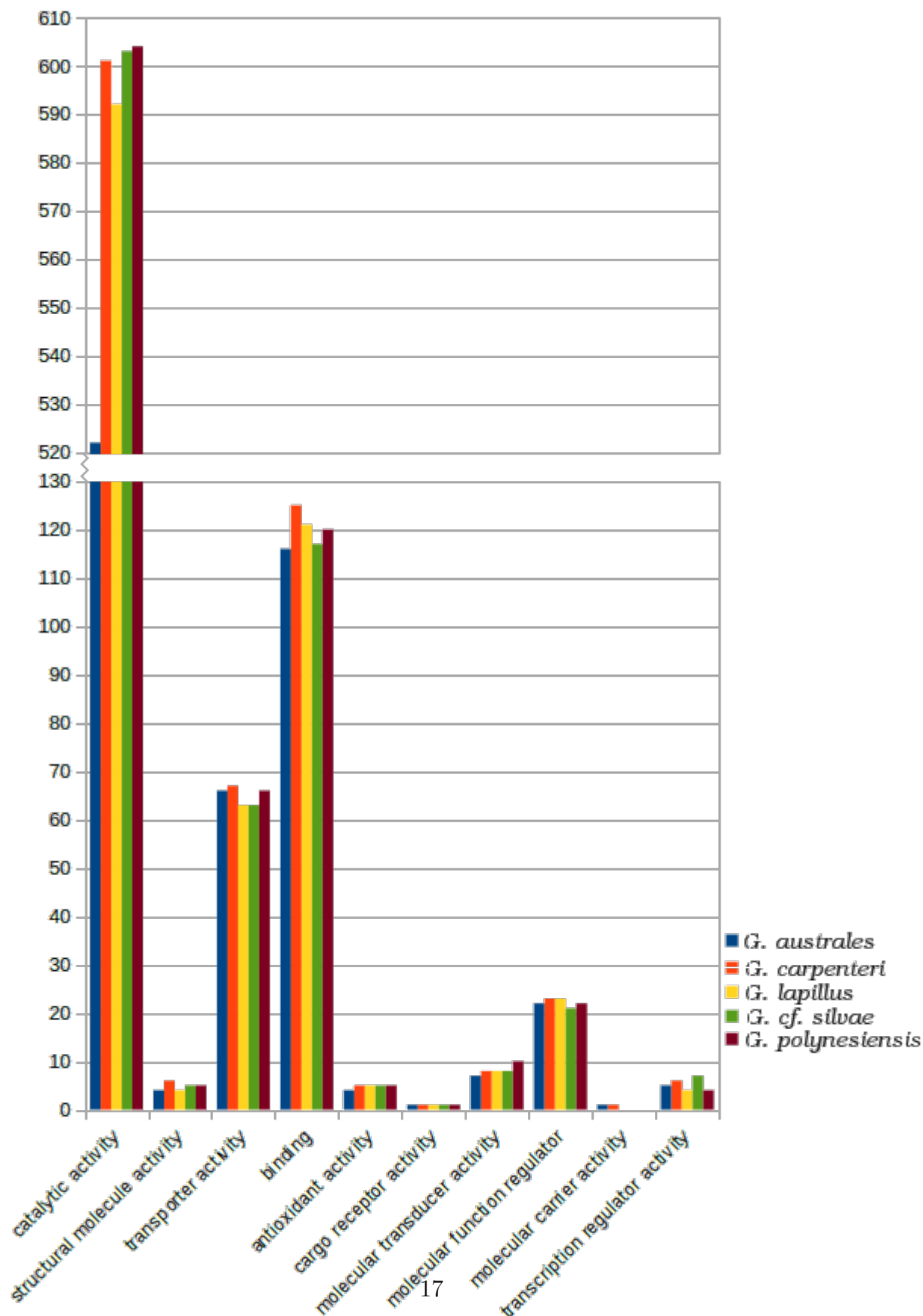


Figure 4: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 1 from Suppl. table 3.

Level 2 GO annotations between *Gambierdiscus* species. At level 2 of GO annotations, difference between species becomes more apparent. While inter-species variations across the molecular, cellular and biological processes (Figs. 11, 12 & 13) are apparent, consistently *G. australes* is underrepresented or absent across all three processes. Conversely, *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, and anatomical structure morphogenesis as well as movement within environment as part of symbiotic interaction in the biological processes. *G. holmesii* had a much higher representation of GO terms matching sperm-egg recognition. The most common molecular process (Fig. 13) mapped to transferase activities, followed by hydrolase activity and oxidoreductase activity. For cellular processes (Fig. 12) 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, in that order, for the biological processes group (Fig. 11).

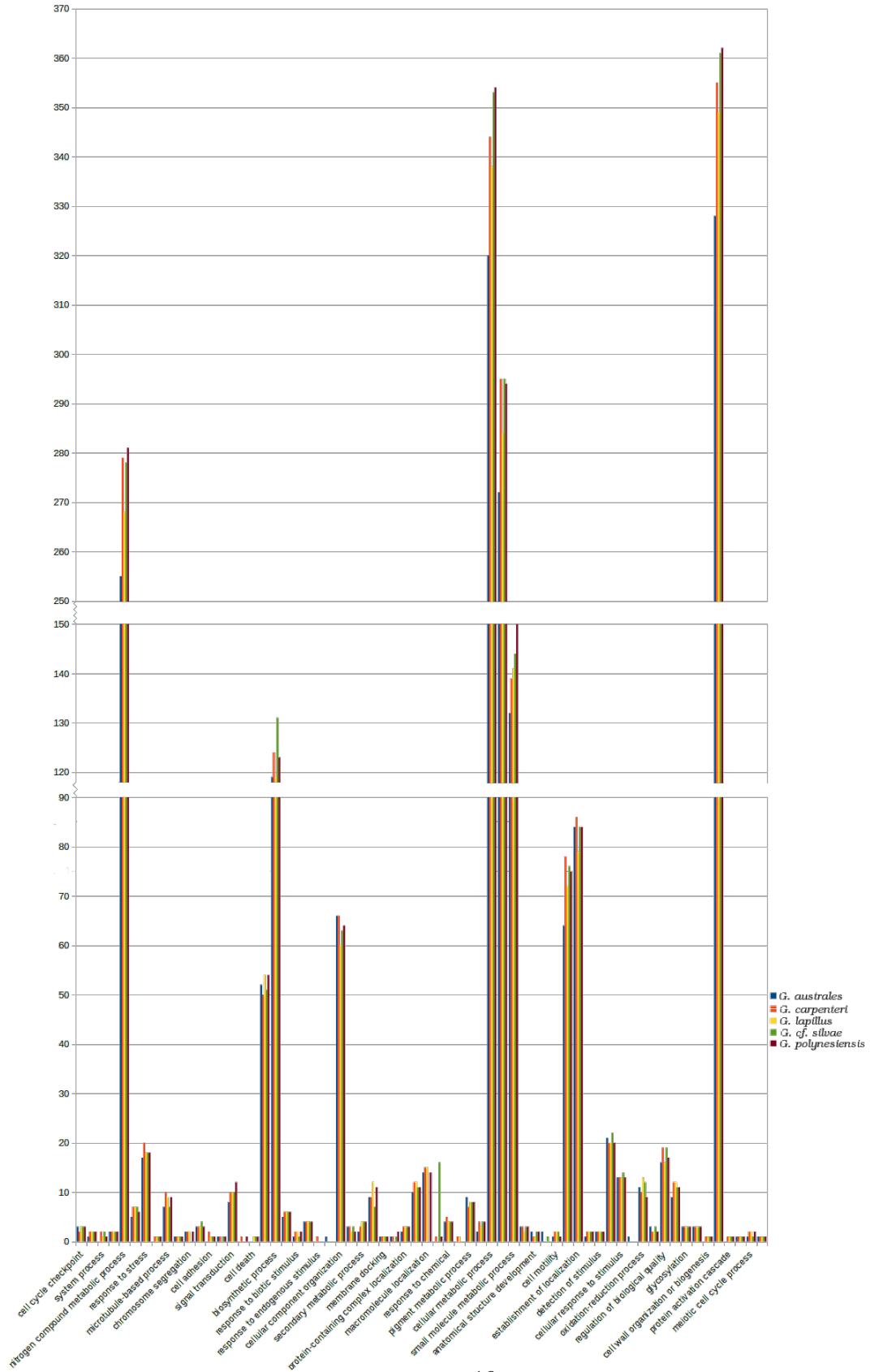


Figure 5: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 2 from Suppl. table 4.

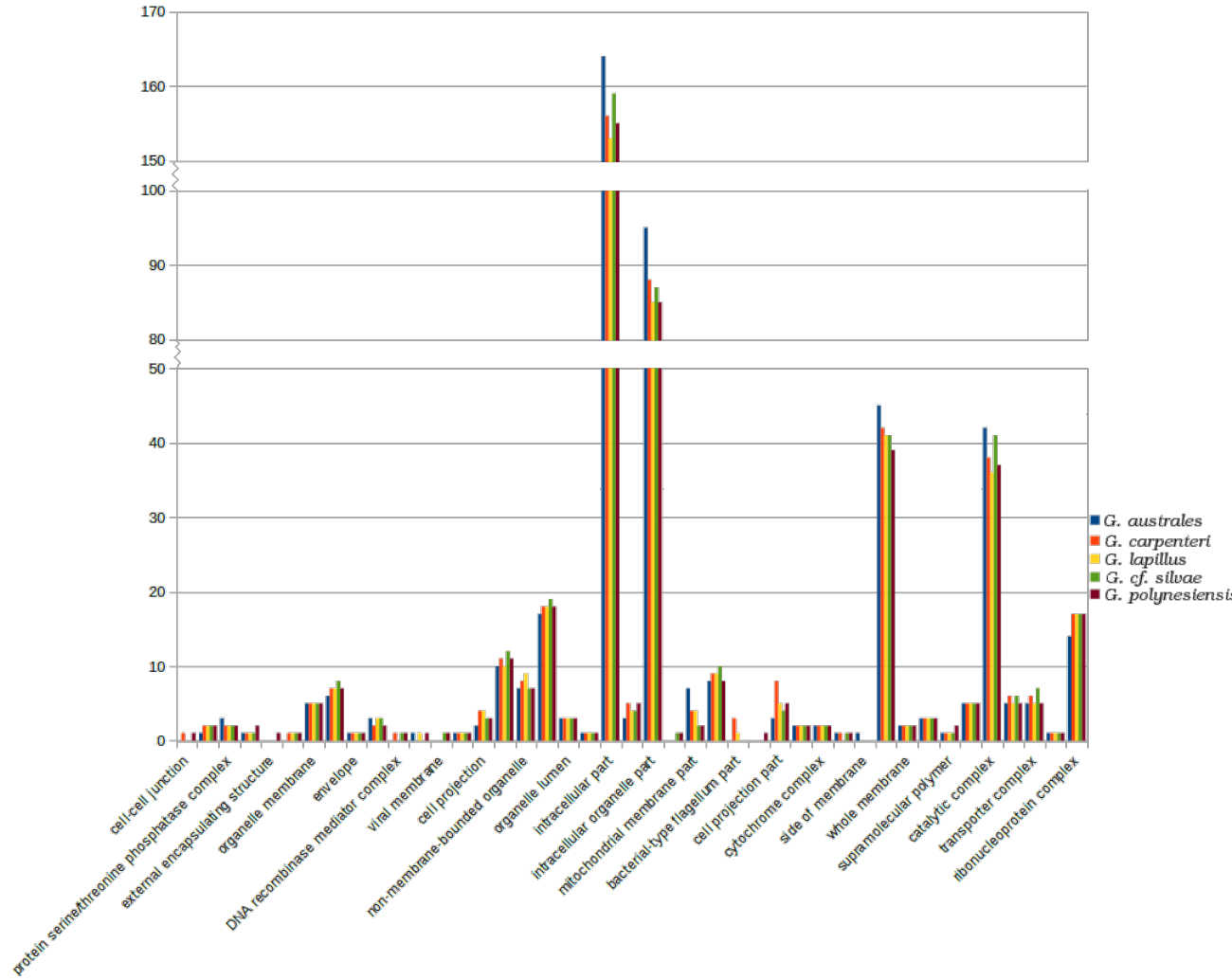


Figure 6: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 2 from Suppl. table 4.

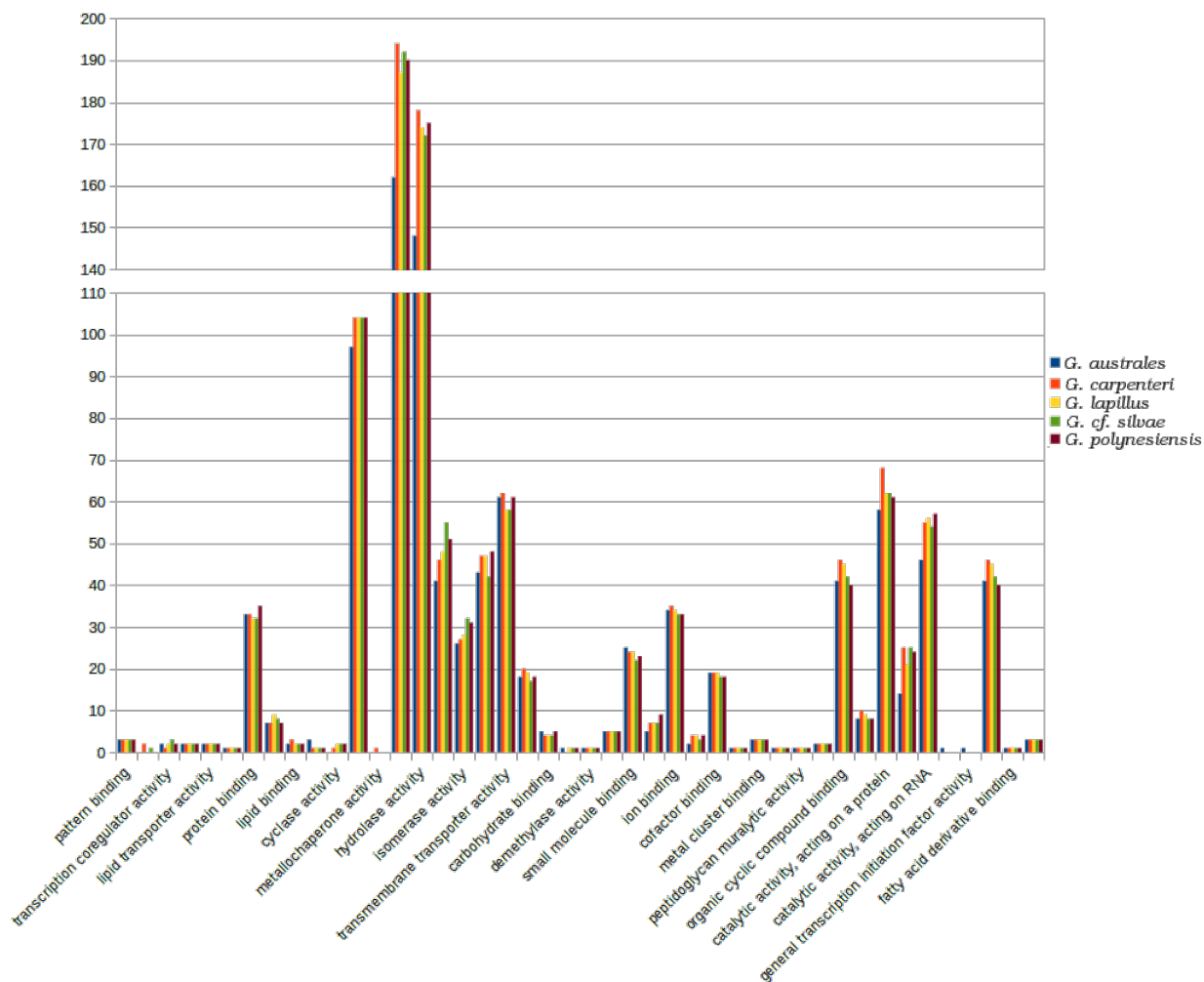


Figure 7: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 2 from Suppl. table 4.

Level 1 GO annotations for pan-transcriptomes. Similarity between the core-, softcore- and unique-transcriptomes 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 synapse 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).

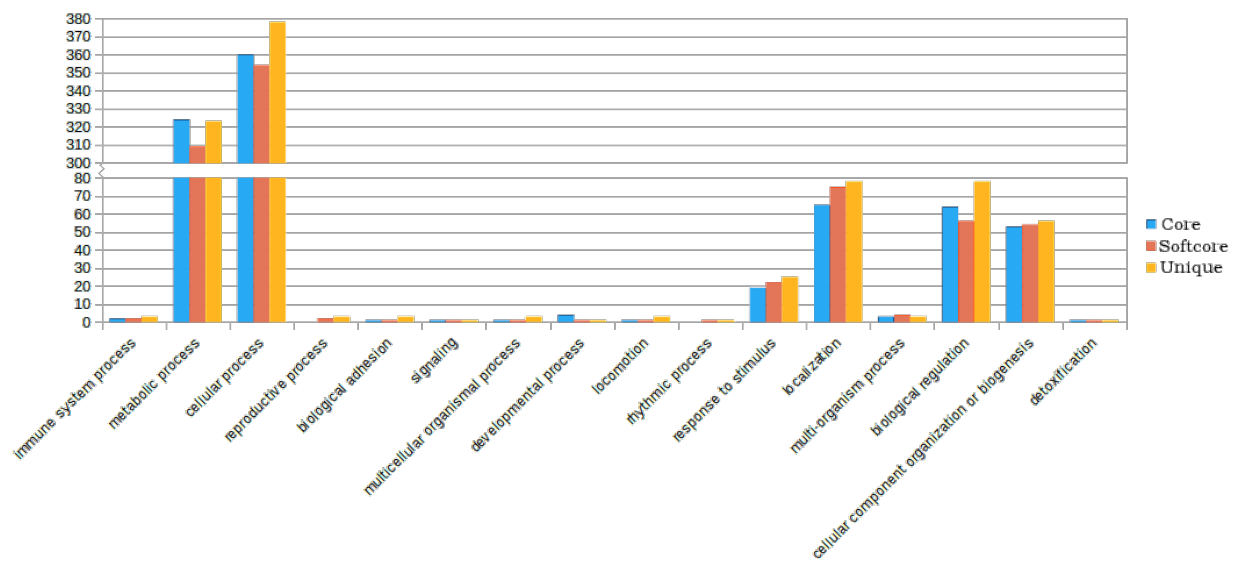


Figure 8: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 1.

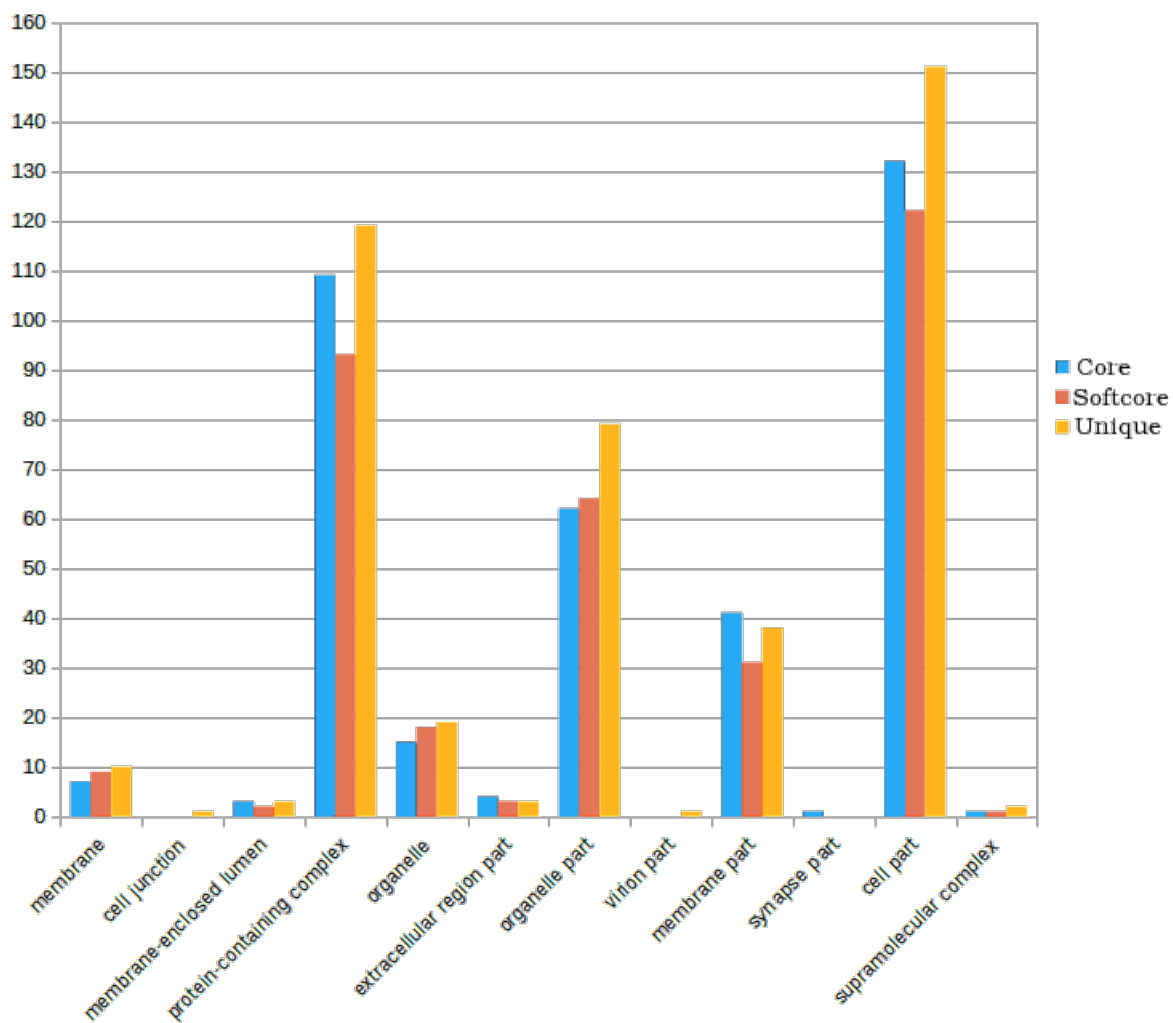


Figure 9: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 1.

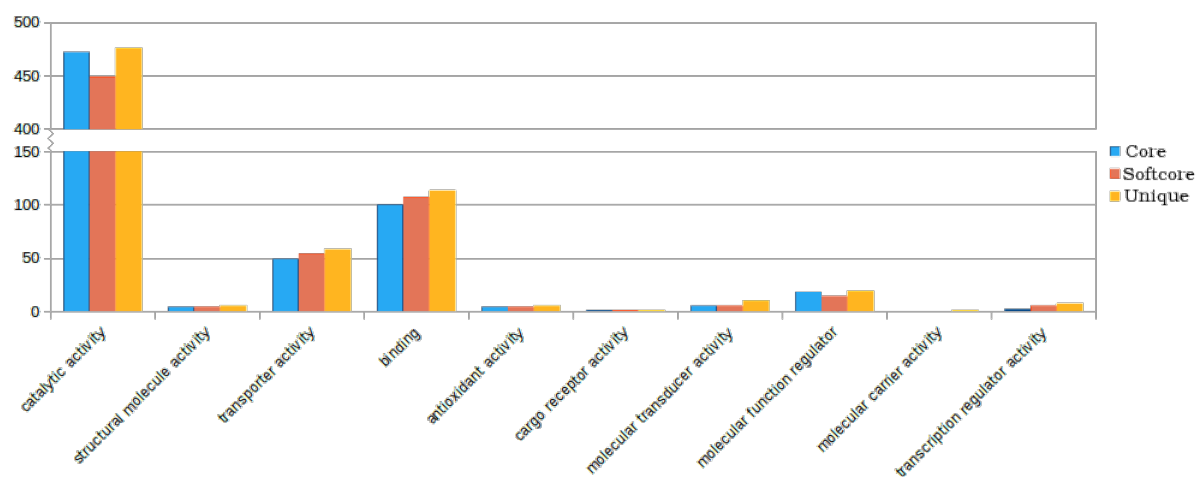


Figure 10: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 1.

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 pan-transcriptome 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, bacteria-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.

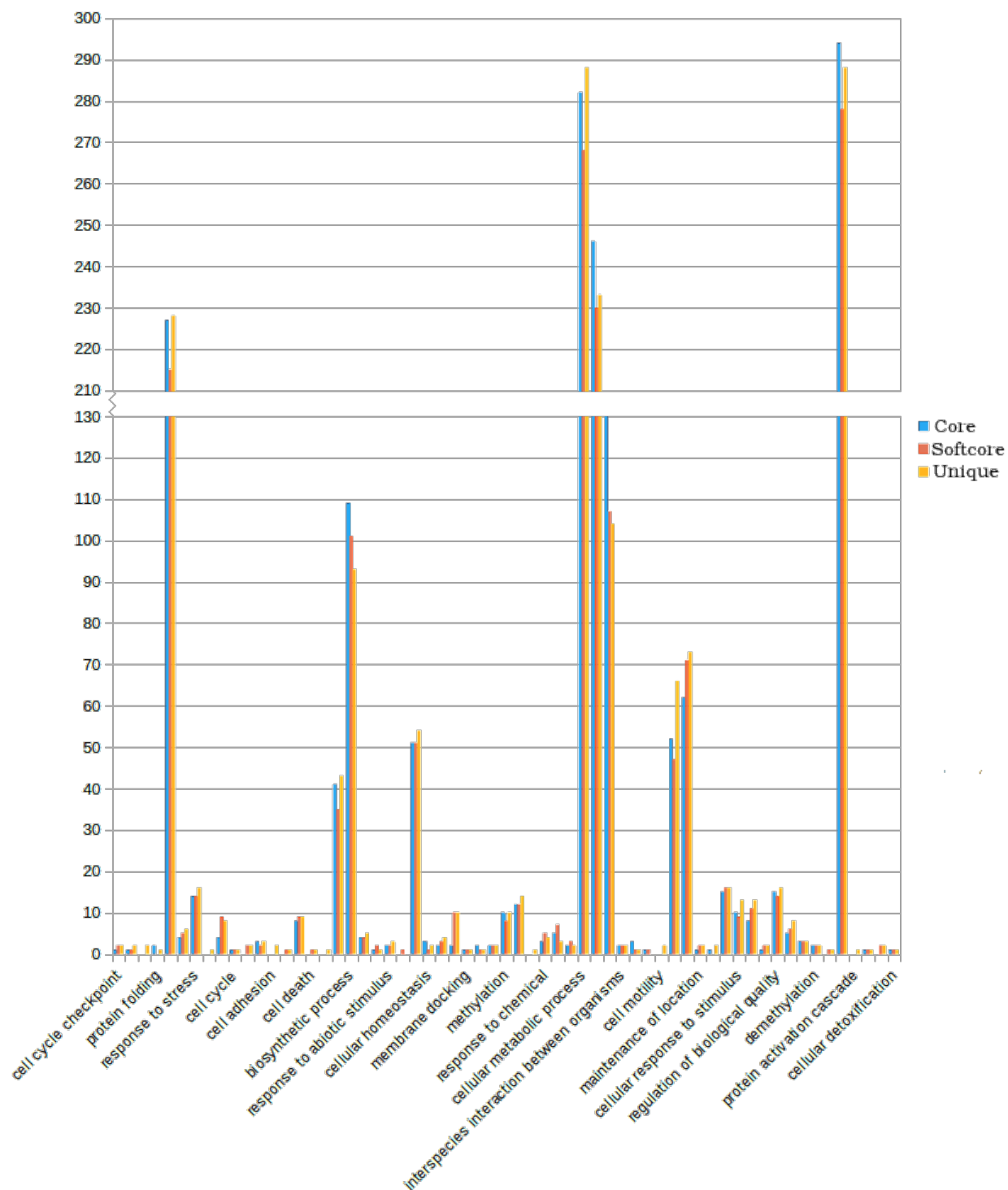


Figure 11: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 2.

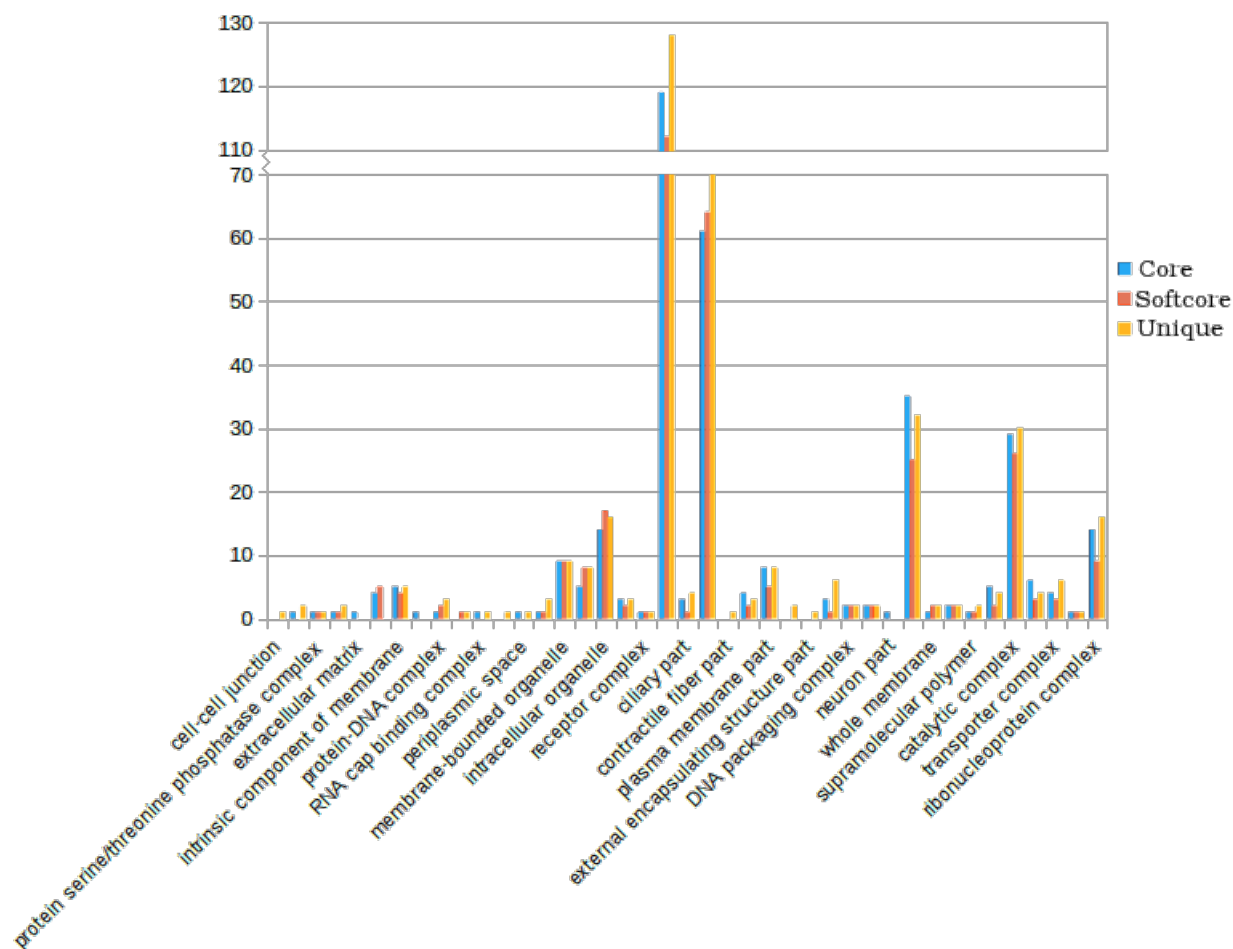


Figure 12: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 2.

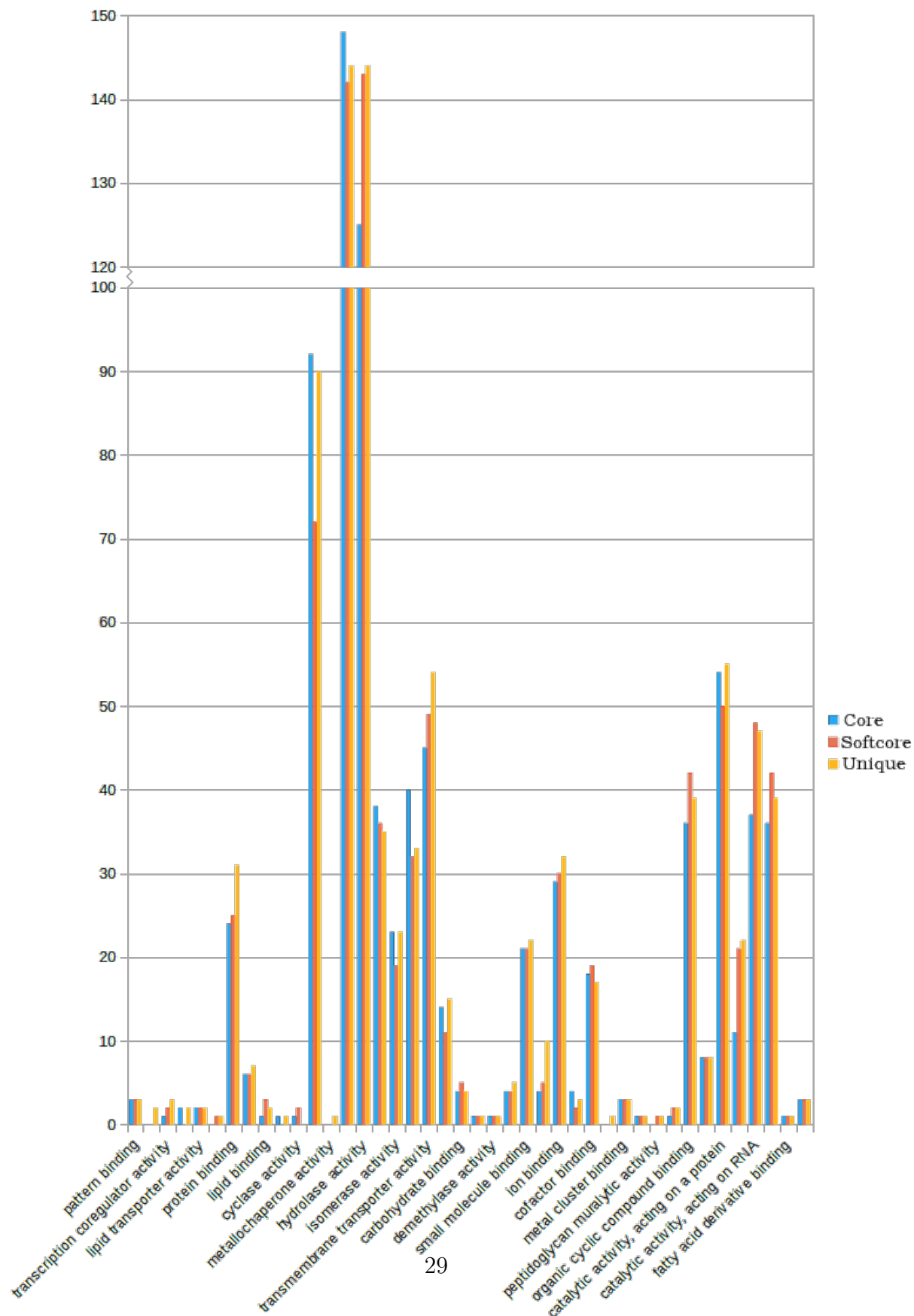


Figure 13: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 2.

4.2 Keto synthase active domain search

A total of 850 contigs were identified with KS domains which assembled into 314 clusters (Fig. 14). Nine clusters contained more than 10 contigs, with the highest number of 130 contigs from all species. 9 clusters contained 10 contigs or more, of which only two did not contain all the taxa examined. 57 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.

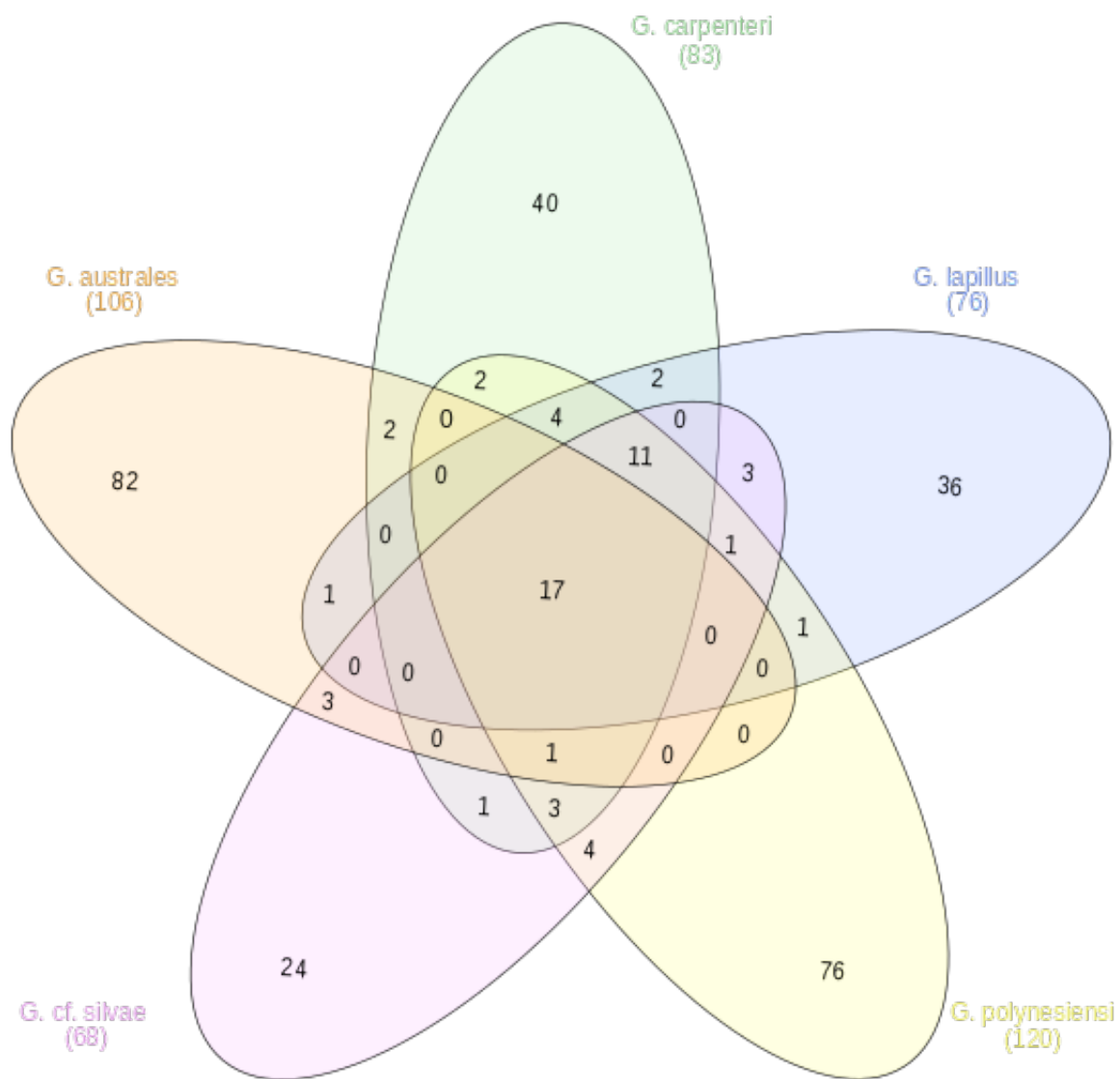


Figure 14: Venn diagram of species in KS clusters.

5 Discussion

Comparing five *Gambierdiscus* species revealed a core, soft-core and unique fraction between the transcriptomes. Further, differences between species with different toxin production characteristics were observed. Interestingly, the abundance of dinoSL was quite low (table 2) compared to the abundance observed by Zhang et al. (2009). Similar to this study, low abundance of spliced leaders has been observed in other dinoflagellates [2, 9]. 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 yet unknown. Interestingly, *G. homesii* and *G. polynesiensis* had the most contigs with dinoSLs. These were from different sequencing runs as well as collected from different geographic locations, but are phylogenetically closely related.

5.1 Core, soft-core and unique 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 soft-core spanned 4 of the 5 taxa. Noticeably, *G. australes* was absent from 86 % of the soft-core dataset which indicates that a large proportion of the soft-core is likely part of the *Gambierdiscus* pan-transcriptome core which was not captured in the *G. australes* sequencing. This is a case where the relevance of a reference pan-transcriptome for sequencing efforts becomes evident.

[25] comprehensive index of genes in *K. brevis* to compare to as well as functional summaries

5.2 core *Gambierdiscus* transcriptome

[25] comprehensive index of genes in *K. brevis* to compare to as well as functional summaries

discuss common & different functions found

Koid 14 pan-transcriptome of 4 prymnesiophyte algae. Compare functional findings (KOG vs. this) and contigs as well as predicted protein coding regions are just a fraction

of the ones here. eg.30,000-56,000 contigs vs. lowest for in this study is 148,972. Other study transcriptomes are part of MMETSP, but even australes here is over 100,000 contigs which is from the same study so more likely it's a Gambi thing rather than a seq thing. same with australes and Koid for peptides predicted, almost double. Way higher for other gambis.

Expression of genes involved in polyketide production

- discuss if different gene sets were expressed between toxic and non- toxic strains (ie. not carp)
- discuss KS containing contigs per species plus distribution and number of contigs in KS clusters
- point at Venn diagram intersections that could be of interest for further investigation for both MTX and CTX
- discuss KS conserved region phylogeny
- **Tim** I'm not sure we know enough about these pathways to do something like fig 5

5.3 Areas for possible improvement in this study

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 bacterial contigs. Hence it would be pertinent to utilize a eukaryote specific RNA-seq strategy for future studies, or 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 []. The importance of including several strains of a species becomes apparent with the discovery of a non-ciguatoxic *G. polynesiensis* strain []. Variability in morphology and toxicity has also been observed in *G. lapillus* [21]. 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

- Usefulness of core transcriptome for RNA sequencing studies
- Investigate poly only KS clusters or clusters with high number of poly reps

Supplementary

- need to add australes

Table 3: GO terms and number of contigs per species at GO ontology level 1.

GO accession	GO terms	<i>G. carpen- teri</i>	<i>G. lapillus</i>	<i>G. polyne- siensis</i>	<i>G. holme- sii</i>
Biological processes					

GO:0002376	immune system process	3	3	3	3
GO:0008152	metabolic process	388	384	396	396
GO:0009987	cellular process	451	440	457	462
GO:0022414	reproductive process	3	2	1	3
GO:0022610	biological adhesion	3	2	2	2
GO:0023052	signaling	1	1	1	1
GO:0032501	multicellular organismal process	3	2	3	2
GO:0032502	developmental process	1	1	3	2
GO:0040007	growth	0	0	1	1
GO:0040011	locomotion	4	4	3	2
GO:0048511	rhythmic process	0	1	1	1
GO:0050896	response to stimulus	30	27	27	27
GO:0051179	localization	90	84	88	88
GO:0051704	multi-organism process	5	4	4	5
GO:0065007	biological regulation	94	85	91	89
GO:0071840	cellular component organization/biogenesis	70	63	67	68
GO:0098754	detoxification	1	1	1	1
Cellular components					
GO:0016020	membrane	9	9	9	10
GO:0030054	cell junction	1	0	0	1
GO:0031974	membrane-enclosed lumen	3	3	3	3
GO:0032991	protein-containing complex	146	141	145	144
GO:0043226	organelle	22	22	22	21
GO:0044421	extracellular region part	4	4	4	4

GO:0044422	organelle part	94	88	89	88
GO:0044423	virion part	1	0	1	1
GO:0044425	membrane part	50	49	50	47
GO:0044456	synapse part	1	0	1	1
GO:0044464	cell part	182	174	178	177
GO:0099080	supramolecular complex	1	1	1	2
Molecular function					
GO:0003824	catalytic activity	601	592	603	604
GO:0005198	structural molecule activity	6	4	5	5
GO:0005215	transporter activity	67	63	63	66
GO:0005488	binding	125	121	117	120
GO:0016209	antioxidant activity	5	5	5	5
GO:0038024	cargo receptor activity	1	1	1	1
GO:0060089	molecular transducer activity	8	8	8	10
GO:0098772	molecular function regulator	23	23	21	22
GO:0140104	molecular carrier activity	1	0	0	0
GO:0140110	transcription regulator activity	6	4	7	4

Table 4: GO terms and number of contigs per species at GO ontology level 2, child terms of Table 3.

GO accession	GO terms	<i>G. carpenteri</i>	<i>G. lapillus</i>	<i>G. polyensis</i>	<i>G. holmesii</i>
Biological processes					
GO:0000075	cell cycle checkpoint	2	3	3	3

GO:0002252	immune effector process	2	2	2	2
GO:0003008	system process	2	1	2	1
GO:0006457	protein folding	2	2	2	2
GO:0006807	nitrogen compound metabolic process	279	268	278	281
GO:0006928	movement of cell or subcellular component	7	7	7	6
GO:0006950	response to stress	20	18	18	18
GO:0006955	immune response	1	1	1	1
GO:0007017	microtubule-based process	10	9	7	9
GO:0007049	cell cycle	1	1	1	1
GO:0007059	chromosome segregation	2	2	0	2
GO:0007154	cell communication	3	3	4	3
GO:0007155	cell adhesion	2	1	1	1
GO:0007163	establishment or maintenance of cell polarity	1	0	1	1
GO:0007165	signal transduction	10	10	10	12
GO:0008037	cell recognition	1	0	0	1
GO:0008219	cell death	0	1	1	1
GO:0009056	catabolic process	50	54	51	54
GO:0009058	biosynthetic process	124	119	131	123
GO:0009605	response to external stimulus	6	6	6	6
GO:0009607	response to biotic stimulus	2	2	1	2

GO:0009628	response to abiotic stimulus	4	4	4	4
GO:0009719	response to endogenous stimulus	1	0	0	0
GO:0016043	cellular component organization	66	60	63	64
GO:0019725	cellular homeostasis	3	2	3	2
GO:0019748	secondary metabolic process	3	4	4	4
GO:0022402	cell cycle process	9	12	7	11
GO:0022406	membrane docking	1	1	1	1
GO:0030029	actin filament-based process	1	0	1	2
GO:0031503	protein-containing complex localization	3	3	3	3
GO:0032259	methylation	12	12	11	11
GO:0033036	macromolecule localization	15	15	0	14
GO:0035036	sperm-egg recognition	1	0	16	1
GO:0042221	response to chemical	5	4	4	4
GO:0042330	taxis	1	1	0	0
GO:0042440	pigment metabolic process	7	8	8	8
GO:0044085	cellular component biogenesis	4	3	4	4
GO:0044237	cellular metabolic process	344	338	353	354
GO:0044238	primary metabolic process	295	284	295	294
GO:0044281	small molecule metabolic process	139	141	144	150

GO:0044419	interspecies inter- action between organisms	3	2	3	3
GO:0048856	anatomical structure development	1	1	2	2
GO:0048869	cellular developmental process	0	0	1	
GO:0048870	cell motility	2	2	2	1
GO:0050789	regulation of biologi- cal process	78	72	76	75
GO:0051234	establishment of local- ization	86	79	84	84
GO:0051235	maintenance of loca- tion	2	2	2	2
GO:0051606	detection of stimulus	2	2	2	2
GO:0051641	cellular localization	20	20	22	20
GO:0051716	cellular response to stimulus	13	13	14	13
GO:0055114	oxidation-reduction process	10	13	12	9
GO:0061919	process utilizing au- tophagic mechanism	2	2	3	2
GO:0065008	regulation of biologi- cal quality	19	16	19	17
GO:0065009	regulation of molecu- lar function	12	12	11	11
GO:0070085	glycosylation	3	3	3	3
GO:0070988	demethylation	3	3	3	3
GO:0071554	cell wall organization or biogenesis	1	1	1	1

GO:0071704	organic substance metabolic process	355	349	361	362
GO:0072376	protein activation cas- cade	1	1	1	1
GO:0140029	exocytic process	1	1	1	1
GO:1903046	meiotic cell cycle pro- cess	2	2	1	2
GO:1990748	cellular detoxification	1	1	1	1
Cellular components					
GO:0005911	cell-cell junction	1	0	0	1
GO:0005929	cilium	2	2	2	2
GO:0008287	protein ser- ine/threonine phos- phatase complex	2	2	2	2
GO:0019867	outer membrane	1	1	1	2
GO:0030312	external encapsulat- ing structure	0	0	0	1
GO:0031012	extracellular matrix	1	1	1	1
GO:0031090	organelle membrane	5	5	5	5
GO:0031224	intrinsic component of membrane	7	7	8	7
GO:0031975	envelope	1	1	1	1
GO:0032993	protein-DNA complex	2	3	3	2
GO:0033061	DNA recombinase me- diator complex	1	0	1	1
GO:0034518	RNA cap binding complex	0	1	0	1
GO:0036338	viral membrane	0	0	1	1
GO:0042597	periplasmic space	1	1	1	1
GO:0042995	cell projection	4	4	3	3
GO:0043227	membrane-bounded organelle	11	10	12	11

GO:0043228	non-membrane-bounded organelle	8	9	7	7
GO:0043229	intracellular organelle	18	18	19	18
GO:0043233	organelle lumen	3	3	3	3
GO:0043235	receptor complex	1	1	1	1
GO:0044424	intracellular part	156	153	159	155
GO:0044441	ciliary part	5	4	4	5
GO:0044446	intracellular organelle part	88	85	87	85
GO:0044449	contractile fiber part	0	0	1	1
GO:0044455	mitochondrial membrane part	4	4	2	2
GO:0044459	plasma membrane part	9	9	10	8
GO:0044461	bacterial-type flagellum part	3	1	0	0
GO:0044462	external encapsulating structure part	0	0	0	1
GO:0044463	cell projection part	8	5	4	5
GO:0044815	DNA packaging complex	2	2	2	2
GO:0070069	cytochrome complex	2	2	2	2
GO:0097458	neuron part	1	0	1	1
GO:0098796	membrane protein complex	42	41	41	39
GO:0098805	whole membrane	2	2	2	2
GO:0099023	tethering complex	3	3	3	3
GO:0099081	supramolecular polymer	1	1	1	2
GO:0120114	Sm-like protein family complex	5	5	5	5

GO:1902494	catalytic complex	38	36	41	37
GO:1990204	oxidoreductase complex	6	5	6	5
GO:1990351	transporter complex	6	5	7	5
GO:1990391	DNA repair complex	1	1	1	1
GO:1990904	ribonucleoprotein complex	17	17	17	17
Molecular function					
GO:0001871	pattern binding	3	3	3	3
GO:0003700	DNA-binding transcription factor activity	2	0	1	0
GO:0003712	transcription coregulator activity	1	2	3	2
GO:0004133	glycogen debranching enzyme activity	2	2	2	2
GO:0005319	lipid transporter activity	2	2	2	2
GO:0005326	neurotransmitter transporter activity	1	1	1	1
GO:0005515	protein binding	33	32	32	35
GO:0008144	drug binding	7	9	8	7
GO:0008289	lipid binding	3	2	2	2
GO:0008565	protein transporter activity	1	1	1	1
GO:0009975	cyclase activity	1	2	2	2
GO:0016491	oxidoreductase activity	104	104	104	104
GO:0016530	metallochaperone activity	1	0	0	0
GO:0016740	transferase activity	194	187	192	190

GO:0016787	hydrolase activity	178	174	172	175
GO:0016829	lyase activity	46	48	55	51
GO:0016853	isomerase activity	27	28	32	31
GO:0016874	ligase activity	47	47	42	48
GO:0022857	transmembrane transporter activity	62	58	58	61
GO:0030234	enzyme regulator activity	20	19	17	18
GO:0030246	carbohydrate binding	4	4	4	5
GO:0030545	receptor regulator activity	0	1	1	1
GO:0032451	demethylase activity	1	1	1	1
GO:0033218	amide binding	5	5	5	5
GO:0036094	small molecule binding	24	24	22	23
GO:0038023	signaling receptor activity	7	7	7	9
GO:0043167	ion binding	35	34	33	33
GO:0044877	protein-containing complex binding	4	4	3	4
GO:0048037	cofactor binding	19	19	18	18
GO:0050824	water binding	1	1	1	1
GO:0051540	metal cluster binding	3	3	3	3
GO:0060090	molecular adaptor activity	1	1	1	1
GO:0061783	peptidoglycan murelytic activity	1	1	1	1
GO:0072341	modified amino acid binding	2	2	2	2
GO:0097159	organic cyclic compound binding	46	45	42	40

GO:0097367	carbohydrate derivative binding	10	9	8	8
GO:0140096	catalytic activity, acting on a protein	68	62	62	61
GO:0140097	catalytic activity, acting on DNA	25	21	25	24
GO:0140098	catalytic activity, acting on RNA	55	56	54	57
GO:1901363	heterocyclic compound binding	46	45	42	40
GO:1901567	fatty acid derivative binding	1	1	1	1
GO:1901681	sulfur compound binding	3	3	3	3

Table 5: GO terms and number of contigs found in core, softcore and pan-transcriptome of *Gambierdiscus* at GO ontology level 1.

GO accession	GO terms	Core	Softcore	Pan
Biological processes				
GO:0002376	immune system process	2	2	3
GO:0008152	metabolic process	324	309	323
GO:0009987	cellular process	360	354	378
GO:0022414	reproductive process	0	2	3
GO:0022610	biological adhesion	1	1	3
GO:0023052	signaling	1	1	1
GO:0032501	multicellular organismal process	1	1	3

GO:0032502	developmental process	4	1	1
GO:0040011	locomotion	1	1	3
GO:0048511	rhythmic process	0	1	1
GO:0050896	response to stimulus	19	22	25
GO:0051179	localization	65	75	78
GO:0051704	multi-organism process	3	4	3
GO:0065007	biological regulation	64	56	78
GO:0071840	cellular component organization or biogenesis	53	54	56
GO:0098754	detoxification	1	1	1
Cellular components				
GO:0016020	membrane	7	9	10
GO:0030054	cell junction	0	0	1
GO:0031974	membrane-enclosed lumen	3	2	3
GO:0032991	protein-containing complex	109	93	119
GO:0043226	organelle	15	18	19
GO:0044421	extracellular region part	4	3	3
GO:0044422	organelle part	62	64	79
GO:0044423	virion part	0	0	1
GO:0044425	membrane part	41	31	38
GO:0044456	synapse part	1	0	0
GO:0044464	cell part	132	122	151
GO:0099080	supramolecular complex	1	1	2
Molecular function				
GO:0003824	catalytic activity	472	449	476

GO:0005198	structural molecule activity	4	4	5
GO:0005215	transporter activity	49	54	58
GO:0005488	binding	100	107	113
GO:0016209	antioxidant activity	4	4	5
GO:0038024	cargo receptor activity	1	1	1
GO:0060089	molecular transducer activity	5	5	10
GO:0098772	molecular function regulator	18	14	19
GO:0140104	molecular carrier activity	0	0	1
GO:0140110	transcription regulator activity	2	5	7

Table 6: GO terms and number of contigs found in core, softcore and pan-transcriptome of *Gambierdiscus* at GO ontology level 2, childer to Table 5.

GO accession	GO terms	Core	Softcore	Pan
Biological processes				
GO:0000075	cell cycle checkpoint	1	2	2
GO:0002252	immune effector process	1	1	2
GO:0003008	system process	0	0	2
GO:0006457	protein folding	2	0	1
GO:0006807	nitrogen compound metabolic process	227	215	228

GO:0006928	movement of cell or subcellular component	4	5	6
GO:0006950	response to stress	14	14	16
GO:0006955	immune response	0	0	1
GO:0007017	microtubule-based process	4	9	8
GO:0007049	cell cycle	1	1	1
GO:0007059	chromosome segregation	0	2	2
GO:0007154	cell communication	3	2	3
GO:0007155	cell adhesion	0	0	2
GO:0007163	establishment or maintenance of cell polarity	0	1	1
GO:0007165	signal transduction	8	9	9
GO:0008037	cell death	0	1	1
GO:0008219	cell death	0	0	1
GO:0009056	catabolic process	41	35	43
GO:0009058	biosynthetic process	109	101	93
GO:0009605	response to external stimulus	4	4	5
GO:0009607	response to biotic stimulus	1	2	1
GO:0009628	response to abiotic stimulus	2	2	3
GO:0009719	response to endogenous stimulus	0	1	0
GO:0016043	cellular component organization	51	51	54
GO:0019725	cellular homeostasis	3	1	2

GO:0019748	secondary metabolic process	2	3	4
GO:0022402	cell cycle process	2	10	10
GO:0022406	membrane docking	1	1	1
GO:0030029	actin filament-based process	2	1	1
GO:0031503	protein-containing complex localization	2	2	2
GO:0032259	methylation	10	8	10
GO:0033036	macromolecule localization	12	12	14
GO:0035036	sperm-egg recognition	0	0	1
GO:0042221	response to chemical	3	5	4
GO:0042440	pigment metabolic process	5	7	3
GO:0044085	cellular component biogenesis	2	3	2
GO:0044237	cellular metabolic process	282	268	288
GO:0044238	primary metabolic process	246	230	233
GO:0044281	small molecule metabolic process	130	107	104
GO:0044419	interspecies interaction between organisms	2	2	2
GO:0048856	anatomical structure development	3	1	1
GO:0048869	cellular developmental process	1	1	0
GO:0048870	cell motility	0	0	2

GO:0050789	regulation of biological process	52	47	66
GO:0051234	establishment of localization	62	71	73
GO:0051235	maintenance of location	1	2	2
GO:0051606	detection of stimulus	1	0	2
GO:0051641	cellular localization	15	16	16
GO:0051716	cellular response to stimulus	10	9	13
GO:0055114	oxidation-reduction process	8	11	13
GO:0061919	process utilizing autophagic mechanism	1	2	2
GO:0065008	regulation of biological quality	15	14	16
GO:0065009	regulation of molecular function	5	6	8
GO:0070085	glycosylation	3	3	3
GO:0070988	demethylation	2	2	2
GO:0071554	cell wall organization or biogenesis	0	1	1
GO:0071704	organic substance metabolic process	294	278	288
GO:0072376	protein activation cascade	0	0	1
GO:0140029	exocytic process	1	1	1
GO:1903046	meiotic cell cycle process	0	2	2
GO:1990748	cellular detoxification	1	1	1
Cellular components				

GO:0005911	cell-cell junction	0	0	1
GO:0005929	cilium	1	0	2
GO:0008287	protein ser- ine/threonine phos- phatase complex	1	1	1
GO:0019867	outer membrane	1	1	2
GO:0031090	extracellular matrix	1	0	0
GO:0031090	organelle membrane	4	5	0
GO:0031224	intrinsic component of membrane	5	4	5
GO:0031975	envelope	1	0	0
GO:0032993	protein-DNA complex	1	2	3
GO:0033061	DNA recombinase me- diator complex	0	1	1
GO:0034518	RNA cap binding complex	1	0	1
GO:0036338	viral membrane	0	0	1
GO:0042597	periplasmic space	1	0	1
GO:0042995	cell projection	1	1	3
GO:0043227	membrane-bounded organelle	9	9	9
GO:0043228	non-membrane- bounded organelle	5	8	8
GO:0043229	intracellular organelle	14	17	16
GO:0043233	organelle lumen	3	2	3
GO:0043235	receptor complex	1	1	1
GO:0044424	intracellular part	119	112	128
GO:0044441	ciliary part	3	1	4
GO:0044446	intracellular organelle part	61	64	74
GO:0044449	contractile fiber part	0	0	1

GO:0044455	mitochondrial mem- brane part	4	2	3
GO:0044459	plasma membrane part	8	5	8
GO:0044461	bacterial-type flagel- lum part	0	0	2
GO:0044462	external encapsulat- ing structure part	0	0	1
GO:0044463	cell projection part	3	1	6
GO:0044815	DNA packaging com- plex	2	2	2
GO:0070069	cytochrome complex	2	2	2
GO:0097458	neuron part	1	0	0
GO:0098796	membrane protein complex	35	25	32
GO:0098805	whole membrane	1	2	2
GO:0099023	tethering complex	2	2	2
GO:0099081	supramolecular poly- mer	1	1	2
GO:0120114	Sm-like protein family complex	5	2	4
GO:1902494	catalytic complex	29	26	30
GO:1990204	oxidoreductase com- plex	6	3	4
GO:1990351	transporter complex	4	3	6
GO:1990391	DNA repair complex	1	1	1
GO:1990904	ribonucleoprotein complex	14	9	16
Molecular function				
GO:0001871	pattern binding	3	3	3

GO:0003700	DNA-binding transcription factor activity	0	0	2
GO:0003712	transcription coregulator activity	1	2	3
GO:0004133	glycogen debranching enzyme activity	2	0	2
GO:0005319	lipid transporter activity	2	2	2
GO:0005326	neurotransmitter transporter activity	0	1	1
GO:0005515	protein binding	24	25	31
GO:0008144	drug binding	6	6	7
GO:0008289	lipid binding	1	3	2
GO:0008565	protein transporter activity	1	0	1
GO:0009975	cyclase activity	1	2	0
GO:0016491	oxidoreductase activity	92	72	90
GO:0016530	metallochaperone activity	0	0	1
GO:0016740	transferase activity	148	142	144
GO:0016787	hydrolase activity	125	143	144
GO:0016829	lyase activity	38	36	35
GO:0016853	isomerase activity	23	19	23
GO:0016874	ligase activity	40	32	33
GO:0022857	transmembrane transporter activity	45	49	54
GO:0030234	enzyme regulator activity	14	11	15
GO:0030246	carbohydrate binding	4	5	4

GO:0030545	receptor regulator activity	1	1	1
GO:0032451	demethylase activity	1	1	1
GO:0033218	amide binding	4	4	5
GO:0036094	small molecule binding	21	21	22
GO:0038023	signaling receptor activity	4	5	10
GO:0043167	ion binding	29	30	32
GO:0044877	protein-containing complex binding	4	2	3
GO:0048037	cofactor binding	18	19	17
GO:0050824	water binding	0	0	1
GO:0051540	metal cluster binding	3	3	3
GO:0060090	molecular adaptor activity	1	1	1
GO:0061783	peptidoglycan murelytic activity	0	1	1
GO:0072341	modified amino acid binding	1	2	2
GO:0097159	organic cyclic compound binding	36	42	39
GO:0097367	carbohydrate derivative binding	8	8	8
GO:0140096	catalytic activity, acting on a protein	54	50	55
GO:0140097	catalytic activity, acting on DNA	11	21	22
GO:0140098	catalytic activity, acting on RNA	37	48	47

GO:1901363	heterocyclic compound binding	36	42	39
GO:1901567	fatty acid derivative binding	1	1	1
GO:1901681	sulfur compound binding	3	3	3

Table 7: KS domains found per cluster and total number of contigs present.

Cluster ID	<i>G. australis</i>	<i>G. carpenteri</i>	<i>G. lapillus</i>	<i>G. polyne-siensis</i>	<i>G. holme-sii</i>	Total contigs
988	6	40	29	24	31	130
8866	3	24	14	24	16	81
3681	7	14	16	9	12	58
1921	3	10	6	4	6	29
46550	3	4	1	8	5	21
215601	0	4	1	8	5	18
360	1	4	3	3	4	15
15645	4	2	0	4	1	11
132980	0	1	4	3	2	10
45086	1	3	1	1	3	9
78009	0	2	2	3	2	9
38915	2	2	2	1	2	9
109763	0	2	0	5	1	8
37859	2	2	1	2	1	8
24847	1	1	1	3	2	8
162333	0	2	2	2	1	7
52333	1	2	1	1	1	6
136782	0	1	2	1	2	6
301971	0	0	2	2	2	6

152898	0	3	1	1	0	5
117472	0	2	1	1	1	5
196360	0	2	1	1	1	5
145445	0	1	1	2	1	5
131919	0	1	0	1	3	5
59207	1	1	1	1	1	5
31669	1	1	1	1	1	5
55678	1	1	1	1	1	5
40462	1	1	1	1	1	5
46899	1	1	1	1	1	5
37886	1	1	1	1	1	5
475329	0	0	0	4	1	5
162320_UTSM0ER9A3_Gambierdiscus- carpenteri_DN15967_c2_g1_i2.p1.faa			0	1	0	4
21082_MMETSP0766_Gambierdiscus- australes_DN32692_c0_g1_i1.p1.faa			0	2	1	4
195242_UTSM0ER9A3_Gambierdiscus- carpenteri_DN17326_c2_g5_i1.p1.faa			1	1	1	4
83891_UTSM0ER9A3_Gambierdiscus- carpenteri_DN13035_c1_g4_i1.p1.faa			1	1	1	4
99486_UTSM0ER9A3_Gambierdiscus- carpenteri_DN13588_c0_g3_i1.p1.faa			1	1	1	4
328911_HG4_Gambierdiscus- lapillus_DN41464_c0_g1_i1.p1.faa			1	3	0	4
643864_HG5_Gambierdiscus- silvae_DN47931_c1_g3_i1.p2.faa			0	0	4	4
186957_UTSM0ER9A3_Gambierdiscus- carpenteri_DN16979_c3_g3_i1.p1.faa			1	1	0	3
193820_UTSM0ER9A3_Gambierdiscus- carpenteri_DN17268_c1_g8_i4.p1.faa			1	1	0	3

147284_UTSM0ER9A3_Gambierdiscus-carpenteri_DN15408.c1_g3_i2.p1.faa	1	1	0	3
116539_UTSM0ER9A3_Gambierdiscus-carpenteri_DN14227.c2_g1_i4.p1.faa	2	0	0	3
242595_UTSM0ER9A3_Gambierdiscus-carpenteri_DN9176.c0_g1_i3.p1.faa	2	0	0	3
524928_CG150Gambierdiscus-polynesiensis_DN43543.c1_g1_i1.p1.faa	0	3	0	3
1040_MMETSIP0766_Gambierdiscus-australes_DN11947.c0_g1_i1.p1.faa	0	0	2	3
38402_MMETSIP0766_Gambierdiscus-australes_DN41494.c1_g1_i3.p1.faa	1	0	0	3
154624_UTSM0ER9A3_Gambierdiscus-carpenteri_DN15679.c0_g6_i1.p1.faa	0	0	0	2
63665_UTSM0ER9A3_Gambierdiscus-carpenteri_DN10182.c0_g1_i2.p1.faa	0	0	0	2
205876_UTSM0ER9A3_Gambierdiscus-carpenteri_DN17803.c0_g4_i1.p1.faa	0	0	0	2
224239_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18618.c3_g6_i1.p1.faa	0	0	0	2
196786_UTSM0ER9A3_Gambierdiscus-carpenteri_DN17387.c2_g2_i1.p1.faa	0	1	0	2
131133_UTSM0ER9A3_Gambierdiscus-carpenteri_DN14782.c2_g4_i3.p1.faa	0	0	1	2
19133_MMETSIP0766_Gambierdiscus-australes_DN30780.c0_g2_i1.p1.faa	0	0	0	2
37007_MMETSIP0766_Gambierdiscus-australes_DN41205.c1_g7_i1.p1.faa	0	0	0	2
424979_CG150Gambierdiscus-polynesiensis_DN34166.c0_g9_i1.p1.faa	0	2	0	2

358554_CG150	Gambierdiscus-	0	0	2	0	2
polynesiensis_DN15070_c0_g1_i1.p2.faa						
408901_CG150	Gambierdiscus-	0	0	2	0	2
polynesiensis_DN32288_c2_g1_i1.p1.faa						
479997_CG150	Gambierdiscus-	0	0	1	1	2
polynesiensis_DN39607_c0_g2_i1.p1.faa						
485470_CG150	Gambierdiscus-	0	0	1	1	2
polynesiensis_DN40097_c0_g1_i2.p1.faa						
258909_HG4_0	Gambierdiscus-	0	0	1	1	2
lapillus_DN22432_c0_g1_i2.p1.faa						
263811_HG4_0	Gambierdiscus-	1	0	0	1	2
lapillus_DN25138_c0_g1_i1.p1.faa						
319034_HG4_0	Gambierdiscus-	1	0	0	1	2
lapillus_DN40675_c3_g1_i2.p1.faa						
319505_HG4_0	Gambierdiscus-	1	0	0	1	2
lapillus_DN40711_c1_g8_i1.p1.faa						
1041_MMETSIP0766	Gambierdiscus-	0	0	0	1	2
australes_DN11947_c0_g2_i1.p1.faa						
27066_MMETSIP0766	Gambierdiscus-	0	0	0	1	2
australes_DN36729_c0_g1_i1.p2.faa						
274389_HG4_0	Gambierdiscus-	2	0	0	0	2
lapillus_DN30113_c0_g1_i2.p1.faa						
46553_MMETSIP0766	Gambierdiscus-	0	0	0	0	2
australes_DN42196_c9_g4_i1.p1.faa						
148669_UTSM0ER9A3	Gambierdiscus-	0	0	0	0	1
carpenteri_DN15462_c1_g7_i1.p1.faa						
234513_UTSM0ER9A3	Gambierdiscus-	0	0	0	0	1
carpenteri_DN23482_c0_g1_i1.p1.faa						
63664_UTSM0ER9A3	Gambierdiscus-	0	0	0	0	1
carpenteri_DN10182_c0_g1_i1.p1.faa						

72166_UTSM0ER9A3_Gambierdiscus-carpenteri_DN1258_c0_g1_i1.p1.faa	0	0	0	1
210660_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18011_c6_g4_i1.p1.faa	0	0	0	1
88291_UTSM0ER9A3_Gambierdiscus-carpenteri_DN13188_c2_g8_i2.p2.faa	0	0	0	1
235070_UTSM0ER9A3_Gambierdiscus-carpenteri_DN25711_c0_g1_i1.p2.faa	0	0	0	1
236919_UTSM0ER9A3_Gambierdiscus-carpenteri_DN33286_c0_g1_i1.p1.faa	0	0	0	1
234708_UTSM0ER9A3_Gambierdiscus-carpenteri_DN24051_c0_g1_i1.p1.faa	0	0	0	1
75892_UTSM0ER9A3_Gambierdiscus-carpenteri_DN12749_c1_g2_i3.p1.faa	0	0	0	1
207498_UTSM0ER9A3_Gambierdiscus-carpenteri_DN17871_c4_g9_i1.p1.faa	0	0	0	1
234298_UTSM0ER9A3_Gambierdiscus-carpenteri_DN22896_c0_g1_i1.p1.faa	0	0	0	1
84448_UTSM0ER9A3_Gambierdiscus-carpenteri_DN13053_c3_g3_i4.p1.faa	0	0	0	1
104611_UTSM0ER9A3_Gambierdiscus-carpenteri_DN13776_c4_g7_i1.p1.faa	0	0	0	1
242597_UTSM0ER9A3_Gambierdiscus-carpenteri_DN9176_c0_g2_i2.p2.faa	0	0	0	1
233698_UTSM0ER9A3_Gambierdiscus-carpenteri_DN2009_c0_g1_i1.p1.faa	0	0	0	1
115505_UTSM0ER9A3_Gambierdiscus-carpenteri_DN14189_c2_g12_i1.p1.faa	0	0	0	1
238946_UTSM0ER9A3_Gambierdiscus-carpenteri_DN4887_c0_g1_i1.p1.faa	0	0	0	1

208524_UTSM0ER9A3_Gambierdiscus-carpenteri_DN17914.c1_g3_i4.p1.faa	0	0	0	1
131131_UTSM0ER9A3_Gambierdiscus-carpenteri_DN14782.c2_g4_i1.p1.faa	0	0	0	1
215621_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18221.c2_g6_i3.p1.faa	0	0	0	1
225926_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18701.c1_g3_i2.p1.faa	0	0	0	1
239297_UTSM0ER9A3_Gambierdiscus-carpenteri_DN5390.c0_g1_i1.p1.faa	0	0	0	1
233616_UTSM0ER9A3_Gambierdiscus-carpenteri_DN19857.c0_g1_i1.p1.faa	0	0	0	1
208525_UTSM0ER9A3_Gambierdiscus-carpenteri_DN17914.c1_g3_i5.p2.faa	0	0	0	1
236171_UTSM0ER9A3_Gambierdiscus-carpenteri_DN30145.c0_g1_i1.p1.faa	0	0	0	1
241217_UTSM0ER9A3_Gambierdiscus-carpenteri_DN7872.c0_g1_i1.p1.faa	0	0	0	1
212813_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18098.c3_g3_i2.p1.faa	0	0	0	1
147705_UTSM0ER9A3_Gambierdiscus-carpenteri_DN15422.c1_g3_i1.p1.faa	0	0	0	1
242594_UTSM0ER9A3_Gambierdiscus-carpenteri_DN9176.c0_g1_i2.p1.faa	0	0	0	1
86631_UTSM0ER9A3_Gambierdiscus-carpenteri_DN13131.c1_g1_i1.p1.faa	0	0	0	1
238247_UTSM0ER9A3_Gambierdiscus-carpenteri_DN38343.c0_g1_i1.p1.faa	0	0	0	1
212812_UTSM0ER9A3_Gambierdiscus-carpenteri_DN18098.c3_g3_i1.p1.faa	0	0	0	1

211703_UTSM0ER9A3_Gambierdiscus-	0	0	0	1
carpenteri_DN18052_c3_g5_i1.p1.faa				
239230_UTSM0ER9A3_Gambierdiscus-	0	0	0	1
carpenteri_DN5288_c0_g1_i1.p1.faa				
103957_UTSM0ER9A3_Gambierdiscus-	0	0	0	1
carpenteri_DN13754_c3_g2_i4.p1.faa				
462243_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN37930_c0_g1_i2.p1.faa				
355979_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN10471_c0_g1_i1.p1.faa				
524904_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN43540_c1_g1_i2.p1.faa				
471036_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN38733_c0_g1_i1.p1.faa				
527904_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN43803_c0_g1_i1.p1.faa				
494332_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN40908_c1_g1_i1.p1.faa				
475327_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN39159_c1_g1_i1.p1.faa				
446377_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN36357_c3_g7_i1.p1.faa				
415511_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN33112_c0_g1_i3.p1.faa				
524930_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN43543_c1_g1_i4.p1.faa				
500254_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN41444_c1_g3_i1.p1.faa				
408903_CG150Gambierdiscus-	0	1	0	1
polynesiensis_DN32288_c3_g1_i1.p1.faa				

211708_UTSMER9A3_Gambierdiscus-carpenteri_DN18052_c3_g5_i7.p1.faa	0	1	0	1
524905_CG150Gambierdiscus-polynesiensis_DN43540_c1_g1_i3.p1.faa	0	1	0	1
528784_CG150Gambierdiscus-polynesiensis_DN47453_c0_g1_i1.p3.faa	0	1	0	1
528223_CG150Gambierdiscus-polynesiensis_DN44935_c0_g1_i1.p2.faa	0	1	0	1
362866_CG150Gambierdiscus-polynesiensis_DN18821_c0_g1_i1.p1.faa	0	1	0	1
408898_CG150Gambierdiscus-polynesiensis_DN32288_c1_g1_i1.p1.faa	0	1	0	1
473656_CG150Gambierdiscus-polynesiensis_DN39000_c2_g2_i1.p1.faa	0	1	0	1
505619_CG150Gambierdiscus-polynesiensis_DN41913_c1_g3_i1.p2.faa	0	1	0	1
357110_CG150Gambierdiscus-polynesiensis_DN13123_c0_g1_i2.p2.faa	0	1	0	1
529123_CG150Gambierdiscus-polynesiensis_DN48937_c0_g1_i1.p1.faa	0	1	0	1
419597_CG150Gambierdiscus-polynesiensis_DN33575_c2_g1_i1.p1.faa	0	1	0	1
486622_CG150Gambierdiscus-polynesiensis_DN40207_c2_g2_i2.p1.faa	0	1	0	1
518712_CG150Gambierdiscus-polynesiensis_DN43045_c0_g2_i6.p1.faa	0	1	0	1
505617_CG150Gambierdiscus-polynesiensis_DN41913_c1_g2_i1.p1.faa	0	1	0	1
419857_CG150Gambierdiscus-polynesiensis_DN33604_c1_g1_i1.p1.faa	0	1	0	1

319033_HG4_Gambierdiscus-lapillus_DN40675_c3_g1_i1.p1.faa	0	1	0	1
505612_CG150Gambierdiscus-polynesiensis_DN41913_c0_g1_i1.p1.faa	0	1	0	1
505621_CG150Gambierdiscus-polynesiensis_DN41913_c1_g5_i1.p2.faa	0	1	0	1
368243_CG150Gambierdiscus-polynesiensis_DN21805_c0_g1_i1.p1.faa	0	1	0	1
531066_CG150Gambierdiscus-polynesiensis_DN7198_c0_g1_i1.p1.faa	0	1	0	1
411779_CG150Gambierdiscus-polynesiensis_DN32643_c5_g2_i3.p2.faa	0	1	0	1
529709_CG150Gambierdiscus-polynesiensis_DN51840_c0_g1_i1.p1.faa	0	1	0	1
424815_CG150Gambierdiscus-polynesiensis_DN34144_c0_g1_i6.p1.faa	0	1	0	1
388829_CG150Gambierdiscus-polynesiensis_DN29147_c0_g1_i1.p1.faa	0	1	0	1
528991_CG150Gambierdiscus-polynesiensis_DN4849_c0_g1_i1.p2.faa	0	1	0	1
529886_CG150Gambierdiscus-polynesiensis_DN52795_c0_g1_i1.p1.faa	0	1	0	1
517572_CG150Gambierdiscus-polynesiensis_DN42942_c0_g1_i1.p1.faa	0	1	0	1
162319_UTSMER9A3_Gambierdiscus-carpenteri_DN15967_c2_g1_i1.p1.faa	0	1	0	1
486374_CG150Gambierdiscus-polynesiensis_DN40177_c0_g2_i3.p1.faa	0	1	0	1
424977_CG150Gambierdiscus-polynesiensis_DN34166_c0_g6_i1.p2.faa	0	1	0	1

480000_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN39607_c0_g2_i4.p1.faa					
524933_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN43543_c1_g1_i7.p1.faa					
529340_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN50363_c0_g1_i1.p1.faa					
382787_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN27509_c0_g1_i1.p1.faa					
455767_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN37290_c0_g4_i1.p1.faa					
454667_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN37192_c1_g3_i1.p1.faa					
505616_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN41913_c1_g1_i3.p1.faa					
408904_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN32288_c3_g2_i1.p1.faa					
519735_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN43127_c3_g5_i1.p1.faa					
524932_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN43543_c1_g1_i6.p1.faa					
419608_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN33575_c2_g2_i1.p1.faa					
489214_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN40447_c0_g1_i2.p1.faa					
407098_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN32057_c0_g1_i2.p1.faa					
486620_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN40207_c2_g1_i2.p2.faa					
529847_CG150	Gambierdiscus-	0	1	0	1
polynesiensis_DN52688_c0_g1_i1.p1.faa					

355910_CG150Gambierdiscus-polynesiensis_DN1036_c0_g1_i1.p2.faa	0	1	0	1
419599_CG150Gambierdiscus-polynesiensis_DN33575_c2_g1_i11.p1.faa	0	1	0	1
368244_CG150Gambierdiscus-polynesiensis_DN21805_c0_g2_i1.p1.faa	0	1	0	1
528301_CG150Gambierdiscus-polynesiensis_DN45312_c0_g1_i1.p1.faa	0	1	0	1
431157_CG150Gambierdiscus-polynesiensis_DN34812_c2_g1_i1.p1.faa	0	1	0	1
429838_CG150Gambierdiscus-polynesiensis_DN3467_c0_g1_i1.p1.faa	0	1	0	1
485799_CG150Gambierdiscus-polynesiensis_DN40132_c0_g3_i1.p1.faa	0	1	0	1
449384_CG150Gambierdiscus-polynesiensis_DN36673_c0_g1_i3.p1.faa	0	1	0	1
530384_CG150Gambierdiscus-polynesiensis_DN55090_c0_g1_i1.p1.faa	0	1	0	1
357109_CG150Gambierdiscus-polynesiensis_DN13123_c0_g1_i1.p2.faa	0	1	0	1
466543_CG150Gambierdiscus-polynesiensis_DN38313_c1_g3_i1.p1.faa	0	1	0	1
367731_CG150Gambierdiscus-polynesiensis_DN21547_c0_g1_i1.p1.faa	0	1	0	1
438506_CG150Gambierdiscus-polynesiensis_DN35575_c0_g1_i7.p1.faa	0	1	0	1
491823_CG150Gambierdiscus-polynesiensis_DN40690_c4_g5_i2.p1.faa	0	1	0	1
530249_CG150Gambierdiscus-polynesiensis_DN54681_c0_g1_i1.p1.faa	0	1	0	1

661643_HG5_ Cambierdiscu silvae_DN57114_c0_g1_i1.p1.faa	0	0	1	1
601478_HG5_ Cambierdiscu silvae_DN43780_c7_g8_i1.p1.faa	0	0	1	1
567939_HG5_ Cambierdiscu silvae_DN35530_c0_g3_i1.p1.faa	0	0	1	1
593688_HG5_ Cambierdiscu silvae_DN42661_c0_g1_i1.p1.faa	0	0	1	1
540524_HG5_ Cambierdiscu silvae_DN20879_c0_g2_i1.p1.faa	0	0	1	1
649671_HG5_ Cambierdiscu silvae_DN48408_c0_g1_i4.p1.faa	0	0	1	1
620146_HG5_ Cambierdiscu silvae_DN45801_c1_g1_i1.p2.faa	0	0	1	1
589550_HG5_ Cambierdiscu silvae_DN41996_c3_g12_i1.p1.faa	0	0	1	1
643868_HG5_ Cambierdiscu silvae_DN47931_c1_g3_i5.p1.faa	0	0	1	1
657026_HG5_ Cambierdiscu silvae_DN48988_c0_g3_i1.p1.faa	0	0	1	1
589562_HG5_ Cambierdiscu silvae_DN41996_c3_g5_i1.p2.faa	0	0	1	1
608846_HG5_ Cambierdiscu silvae_DN44648_c2_g1_i1.p1.faa	0	0	1	1
593690_HG5_ Cambierdiscu silvae_DN42661_c0_g2_i3.p1.faa	0	0	1	1
550256_HG5_ Cambierdiscu silvae_DN27602_c0_g2_i1.p1.faa	0	0	1	1
608853_HG5_ Cambierdiscu silvae_DN44648_c2_g6_i1.p1.faa	0	0	1	1

559711_HG5_ Cambierdiscus silvae_DN32102_c0_g1_i2.p1.faa	0	0	1	1
575231_HG5_ Cambierdiscus silvae_DN38322_c1_g2_i1.p1.faa	0	0	1	1
591087_HG5_ Cambierdiscus silvae_DN42232_c1_g4_i1.p2.faa	0	0	1	1
657027_HG5_ Cambierdiscus silvae_DN48988_c0_g3_i2.p1.faa	0	0	1	1
540525_HG5_ Cambierdiscus silvae_DN20879_c0_g3_i1.p1.faa	0	0	1	1
601479_HG5_ Cambierdiscus silvae_DN43780_c7_g9_i1.p1.faa	0	0	1	1
589619_HG5_ Cambierdiscus silvae_DN42009_c0_g1_i3.p1.faa	0	0	1	1
596728_HG5_ Cambierdiscus silvae_DN43120_c1_g4_i4.p1.faa	0	0	1	1
254977_HG4_ Cambierdiscus lapillus_DN19871_c0_g1_i1.p1.faa	1	0	0	1
244474_HG4_ Cambierdiscus lapillus_DN10661_c0_g1_i1.p1.faa	1	0	0	1
354441_HG4_ Cambierdiscus lapillus_DN7536_c0_g1_i1.p2.faa	1	0	0	1
277633_HG4_ Cambierdiscus lapillus_DN31491_c0_g2_i1.p1.faa	1	0	0	1
312699_HG4_ Cambierdiscus lapillus_DN40082_c0_g1_i1.p1.faa	1	0	0	1
319501_HG4_ Cambierdiscus lapillus_DN40711_c1_g5_i1.p1.faa	1	0	0	1
244476_HG4_ Cambierdiscus lapillus_DN10661_c0_g2_i1.p1.faa	1	0	0	1

355588_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN9793_c0_g1_i1.p1.faa						
351360_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN46619_c0_g1_i1.p2.faa						
319490_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN40711_c1_g10_i1.p1.faa						
249529_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN15767_c0_g4_i1.p1.faa						
350445_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN4403_c0_g2_i1.p1.faa						
249527_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN15767_c0_g2_i1.p1.faa						
247959_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN14263_c0_g2_i1.p1.faa						
354628_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN8017_c0_g2_i1.p1.faa						
327310_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN41349_c0_g1_i2.p1.faa						
245201_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN11411_c0_g1_i1.p1.faa						
328839_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN41459_c1_g5_i1.p1.faa						
332373_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN41718_c2_g1_i1.p1.faa						
310068_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN39797_c2_g1_i1.p1.faa						
355491_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN9601_c0_g2_i1.p1.faa						
264742_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN25642_c0_g1_i1.p1.faa						

310329_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN39821_c0_g4_i1.p1.faa						
351275_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN46352_c0_g1_i1.p1.faa						
298246_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN38038_c1_g5_i1.p1.faa						
312700_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN40082_c0_g2_i1.p1.faa						
245202_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN11411_c0_g1_i2.p1.faa						
270811_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN28598_c0_g3_i1.p1.faa						
311957_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN40004_c3_g1_i2.p1.faa						
270397_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN2840_c0_g1_i2.p1.faa						
351199_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN46156_c0_g1_i1.p1.faa						
308197_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN39584_c0_g1_i1.p1.faa						
354586_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN792_c0_g1_i1.p1.faa						
350059_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN43172_c0_g1_i1.p1.faa						
264744_HG4_	Cambierdiscus		1	0	0	1
lapillus_DN25642_c0_g2_i1.p1.faa						
27277_MMETSP0766_Gambierdiscus-			0	0	0	1
australes_DN36847_c0_g3_i1.p1.faa						
38401_MMETSP0766_Gambierdiscus-			0	0	0	1
australes_DN41494_c1_g1_i2.p1.faa						

14801_MMETSP0766_Gambierdiscus-australes_DN27057_c0_g3_i1.p1.faa	0	0	0	1
38397_MMETSP0766_Gambierdiscus-australes_DN41494_c0_g2_i1.p1.faa	0	0	0	1
19134_MMETSP0766_Gambierdiscus-australes_DN30780_c0_g3_i1.p1.faa	0	0	0	1
33030_MMETSP0766_Gambierdiscus-australes_DN39895_c0_g4_i1.p1.faa	0	0	0	1
35041_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g8_i1.p2.faa	0	0	0	1
40251_MMETSP0766_Gambierdiscus-australes_DN41766_c4_g24_i2.p1.faa	0	0	0	1
18687_MMETSP0766_Gambierdiscus-australes_DN30415_c0_g4_i1.p1.faa	0	0	0	1
18486_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g2_i1.p1.faa	0	0	0	1
61098_MMETSP0766_Gambierdiscus-australes_DN6084_c0_g1_i1.p1.faa	0	0	0	1
36998_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g1_i1.p1.faa	0	0	0	1
13967_MMETSP0766_Gambierdiscus-australes_DN26272_c0_g1_i1.p1.faa	0	0	0	1
19163_MMETSP0766_Gambierdiscus-australes_DN30800_c0_g7_i1.p2.faa	0	0	0	1
35033_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g1_i1.p1.faa	0	0	0	1
18548_MMETSP0766_Gambierdiscus-australes_DN30296_c0_g2_i1.p1.faa	0	0	0	1
36992_MMETSP0766_Gambierdiscus-australes_DN41205_c0_g1_i1.p1.faa	0	0	0	1

36641_MMETSP0766_Gambierdiscus-australes_DN41111_c0_g2_i2.p1.faa	0	0	0	1
37002_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g4_i1.p1.faa	0	0	0	1
28106_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g2_i1.p1.faa	0	0	0	1
72_MMETSP0766_Gambierdiscus-australes_DN10092_c0_g1_i1.p2.faa	0	0	0	1
15646_MMETSP0766_Gambierdiscus-australes_DN27800_c0_g3_i1.p1.faa	0	0	0	1
27067_MMETSP0766_Gambierdiscus-australes_DN36729_c0_g2_i1.p1.faa	0	0	0	1
24849_MMETSP0766_Gambierdiscus-australes_DN35413_c0_g3_i1.p1.faa	0	0	0	1
35035_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g3_i1.p1.faa	0	0	0	1
35352_MMETSP0766_Gambierdiscus-australes_DN40756_c0_g1_i1.p1.faa	0	0	0	1
13100_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g3_i1.p2.faa	0	0	0	1
38396_MMETSP0766_Gambierdiscus-australes_DN41494_c0_g1_i1.p1.faa	0	0	0	1
27068_MMETSP0766_Gambierdiscus-australes_DN36729_c0_g3_i1.p1.faa	0	0	0	1
35359_MMETSP0766_Gambierdiscus-australes_DN40756_c1_g4_i1.p1.faa	0	0	0	1
30964_MMETSP0766_Gambierdiscus-australes_DN38922_c0_g1_i1.p1.faa	0	0	0	1
30406_MMETSP0766_Gambierdiscus-australes_DN38631_c0_g4_i1.p1.faa	0	0	0	1

36994_MMETSP0766_Gambierdiscus-australes_DN41205_c0_g3_i1.p1.faa	0	0	0	1
13103_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g6_i1.p1.faa	0	0	0	1
18485_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g1_i1.p1.faa	0	0	0	1
23610_MMETSP0766_Gambierdiscus-australes_DN34624_c0_g3_i1.p2.faa	0	0	0	1
18487_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g2_i2.p1.faa	0	0	0	1
46548_MMETSP0766_Gambierdiscus-australes_DN42196_c9_g10_i1.p2.faa	0	0	0	1
15765_MMETSP0766_Gambierdiscus-australes_DN27921_c0_g3_i1.p2.faa	0	0	0	1
30418_MMETSP0766_Gambierdiscus-australes_DN38640_c0_g5_i1.p2.faa	0	0	0	1
17953_MMETSP0766_Gambierdiscus-australes_DN29796_c0_g1_i1.p1.faa	0	0	0	1
37006_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g6_i1.p2.faa	0	0	0	1
13101_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g4_i1.p1.faa	0	0	0	1
46559_MMETSP0766_Gambierdiscus-australes_DN42196_c9_g6_i3.p1.faa	0	0	0	1
17396_MMETSP0766_Gambierdiscus-australes_DN2924_c0_g1_i1.p1.faa	0	0	0	1
18547_MMETSP0766_Gambierdiscus-australes_DN30296_c0_g1_i1.p1.faa	0	0	0	1
7245_MMETSP0766_Gambierdiscus-australes_DN20901_c0_g2_i1.p1.faa	0	0	0	1

17216_MMETSP0766_Gambierdiscus-australes_DN29099_c0_g2_i1.p1.faa	0	0	0	1
30404_MMETSP0766_Gambierdiscus-australes_DN38631_c0_g1_i1.p1.faa	0	0	0	1
40250_MMETSP0766_Gambierdiscus-australes_DN41766_c4_g23_i1.p2.faa	0	0	0	1
41420_MMETSP0766_Gambierdiscus-australes_DN41862_c2_g4_i1.p1.faa	0	0	0	1
26875_MMETSP0766_Gambierdiscus-australes_DN36626_c0_g1_i1.p1.faa	0	0	0	1
28107_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g4_i1.p1.faa	0	0	0	1
14799_MMETSP0766_Gambierdiscus-australes_DN27057_c0_g1_i1.p1.faa	0	0	0	1
37863_MMETSP0766_Gambierdiscus-australes_DN41388_c0_g2_i1.p1.faa	0	0	0	1
38400_MMETSP0766_Gambierdiscus-australes_DN41494_c1_g1_i1.p1.faa	0	0	0	1
23609_MMETSP0766_Gambierdiscus-australes_DN34624_c0_g2_i1.p2.faa	0	0	0	1
10084_MMETSP0766_Gambierdiscus-australes_DN23190_c0_g1_i1.p1.faa	0	0	0	1
17955_MMETSP0766_Gambierdiscus-australes_DN29796_c0_g3_i1.p1.faa	0	0	0	1
13099_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g2_i1.p2.faa	0	0	0	1
37003_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g4_i2.p1.faa	0	0	0	1
42718_MMETSP0766_Gambierdiscus-australes_DN41959_c8_g1_i1.p1.faa	0	0	0	1

52639_MMETSP0766_Gambierdiscus-australes_DN45380.c0_g1_i1.p1.faa	0	0	0	1
38398_MMETSP0766_Gambierdiscus-australes_DN41494.c0_g3_i1.p2.faa	0	0	0	1
28109_MMETSP0766_Gambierdiscus-australes_DN37278.c0_g6_i1.p1.faa	0	0	0	1
24850_MMETSP0766_Gambierdiscus-australes_DN35413.c0_g4_i1.p1.faa	0	0	0	1
35358_MMETSP0766_Gambierdiscus-australes_DN40756.c1_g1_i1.p1.faa	0	0	0	1
15764_MMETSP0766_Gambierdiscus-australes_DN27921.c0_g2_i1.p1.faa	0	0	0	1
28108_MMETSP0766_Gambierdiscus-australes_DN37278.c0_g5_i1.p1.faa	0	0	0	1
18688_MMETSP0766_Gambierdiscus-australes_DN30415.c0_g5_i1.p1.faa	0	0	0	1
35357_MMETSP0766_Gambierdiscus-australes_DN40756.c0_g5_i1.p1.faa	0	0	0	1
35353_MMETSP0766_Gambierdiscus-australes_DN40756.c0_g1_i3.p1.faa	0	0	0	1
37009_MMETSP0766_Gambierdiscus-australes_DN41205.c1_g9_i1.p1.faa	0	0	0	1
15763_MMETSP0766_Gambierdiscus-australes_DN27921.c0_g1_i1.p1.faa	0	0	0	1
41417_MMETSP0766_Gambierdiscus-australes_DN41862.c2_g2_i1.p1.faa	0	0	0	1
28600_MMETSP0766_Gambierdiscus-australes_DN37530.c0_g2_i1.p1.faa	0	0	0	1
7400_MMETSP0766_Gambierdiscus-australes_DN21004.c0_g1_i1.p1.faa	0	0	0	1

35356_MME	TSP0766_Gam	Bierdiscus-	0	0	0	1
australes_DN	40756_c0_g4_i1.p1.faa					
30416_MME	TSP0766_Gam	Bierdiscus-	0	0	0	1
australes_DN	38640_c0_g3_i1.p1.faa					
18686_MME	TSP0766_Gam	Bierdiscus-	0	0	0	1
australes_DN	30415_c0_g2_i1.p1.faa					
16842_MME	TSP0766_Gam	Bierdiscus-	0	0	0	1
australes_DN	28731_c0_g3_i1.p2.faa					

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] CONSORTIUM, G. O. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research* 45, D1 (2016), D331–D338.
- [5] CONSORTIUM, U. Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* 39, suppl.1 (2010), D214–D219.
- [6] EDDY, S., AND WHEELER, T. HMMER: biosequence analysis using profile hidden Markov models, 2015. hmmer.org/.
- [7] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32, 5 (2004), 1792–1797.
- [8] 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.
- [9] 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.
- [10] HAAS, B., AND PAPANICOLAOU, A. TransDecoder (find coding regions within transcripts), 2016.

- [11] 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.
- [12] HE, F., AND MASLOV, S. Pan-and core-network analysis of co-expression genes in a model plant. *Scientific reports* 6 (2016), 38956.
- [13] 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.
- [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] KAHLKE, T., AND RALPH, P. J. Basta—taxonomic classification of sequences and sequence bins using Last Common Ancestor estimations. *Methods in Ecology and Evolution* (2018).
- [17] 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).
- [18] 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).
- [19] 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.
- [20] 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.
 - [21] 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.
 - [22] LAPIERRE, P., AND GOGARTEN, J. P. Estimating the size of the bacterial pan-genome. *Trends in genetics* 25, 3 (2009), 107–110.
 - [23] 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.
 - [24] 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.
 - [25] LIDIE, K. B., RYAN, J. C., BARBIER, M., AND VAN DOLAH, F. M. Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of DNA microarray. *Marine Biotechnology* 7, 5 (2005), 481–493.
 - [26] MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V., AND RAPPUOLI, R. The microbial pan-genome. *Current opinion in genetics & development* 15, 6 (2005), 589–594.
 - [27] 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 polyketide synthase genes. *BMC genomics* 16, 1 (2015), 27.
- [28] MOUSTAFA, A., EVANS, A. N., KULIS, D. M., HACKETT, J. D., ERDNER, D. L., ANDERSON, D. M., AND BHATTACHARYA, D. Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One* 5, 3 (2010), e9688.
 - [29] 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.
 - [30] 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.
 - [31] 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.
 - [32] 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.
 - [33] 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.
 - [34] 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.

- [35] 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 *Emiliana* underpins its global distribution. *Nature* 499, 7457 (2013), 209.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 9 (2014), 1312–1313.
- [41] 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.
- [42] VERNIKOS, G., MEDINI, D., RILEY, D. R., AND TETTELIN, H. Ten years of pan-genome analyses. *Current opinion in microbiology* 23 (2015), 148–154.
- [43] 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.

- [44] ZHANG, H., AND LIN, S. Retrieval of missing spliced leader in dinoflagellates. *PLoS One* 4, 1 (2009), e4129.