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 anaologs 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* cf. *silvae* 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 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 gonyalacales 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 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 MMETSP 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 reveled 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	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*
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 use 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).

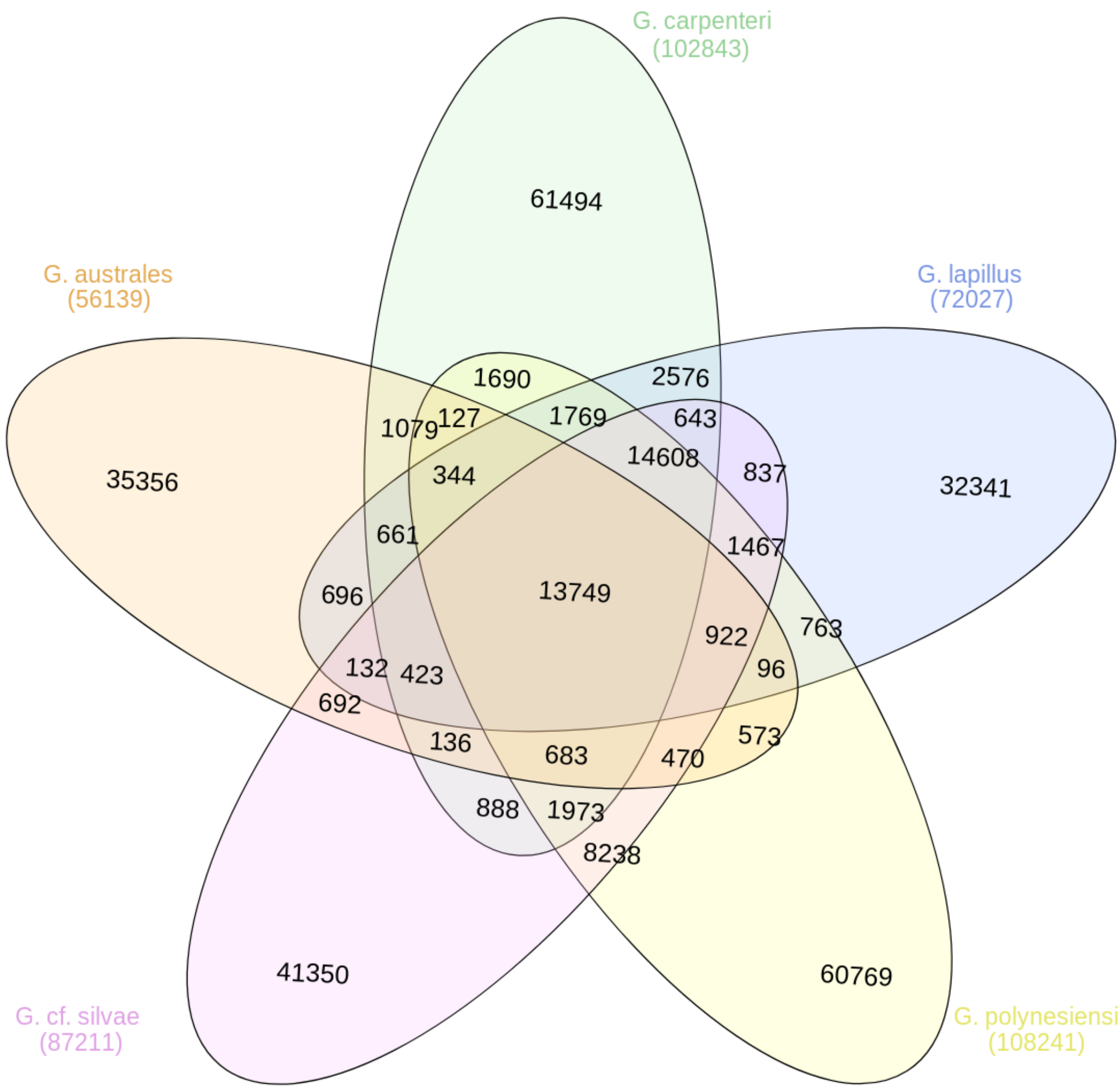

Figure 1: Venn diagram of species distribution across clusters.

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

Species	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*
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

Tim: It seems kinda conspicuous that the unique clusters of *G. carpenteri* & *G. poly* are almost twice the number of *G. lapillus* and *G. silvae*, 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 Comparison of *Gambierdiscus* inter-species transcriptome annotations

The GOs were split up into the three functional groups defined by the consortium: 1) Molecular processes (Figs. 4 & 7) defined as biochemical or a macromolecule directly interacting with other molecules; 2) Cellular components (Figs. 2 & 6) defined by the location within the cell where a molecular process takes place; and 3) Biological process (Figs 3 & 5) 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 hirachical 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). 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).
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 metobolic 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 (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 lease represented biological GO annotation was related to growth with only one annotation for *G. holmesii* and *G. polynesiensis*.

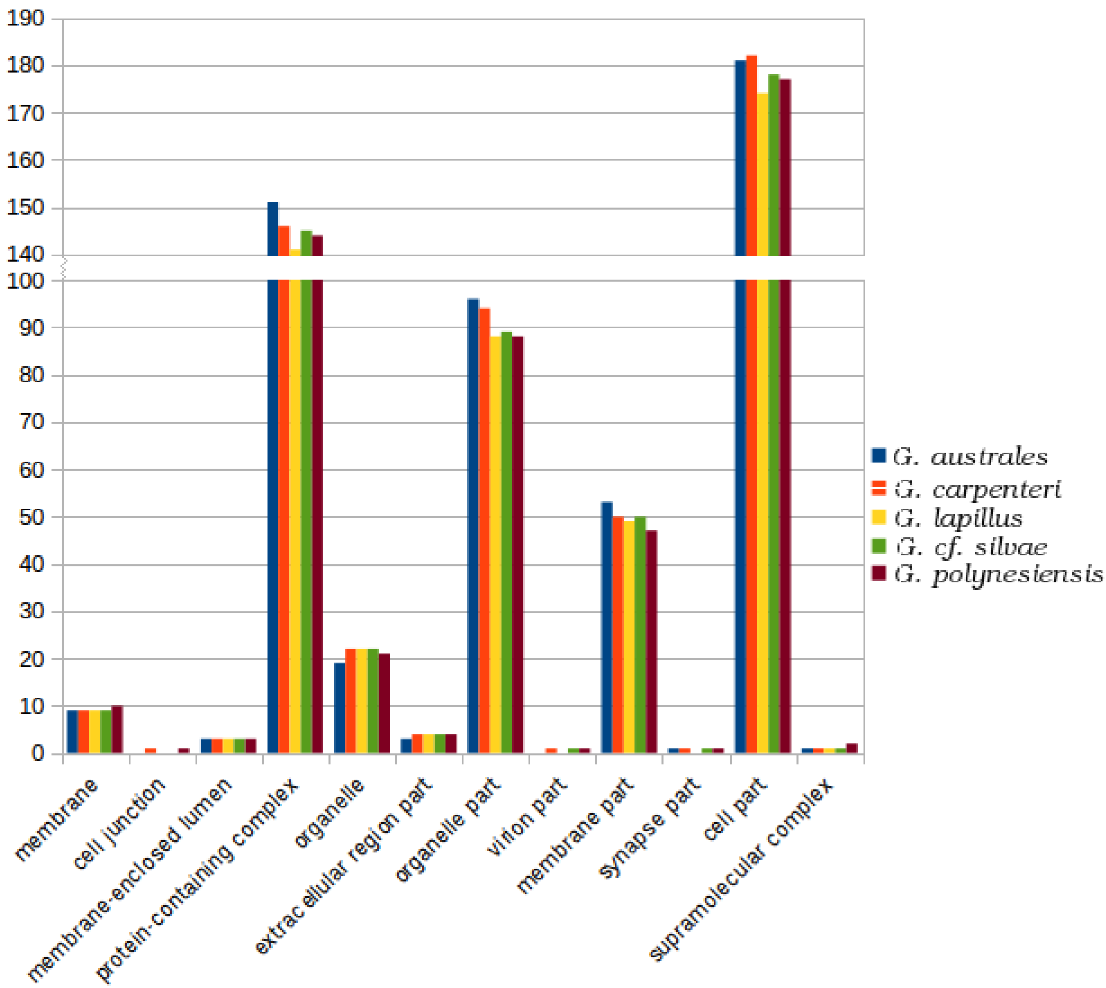

Figure 2: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 1 from Suppl. table 6.

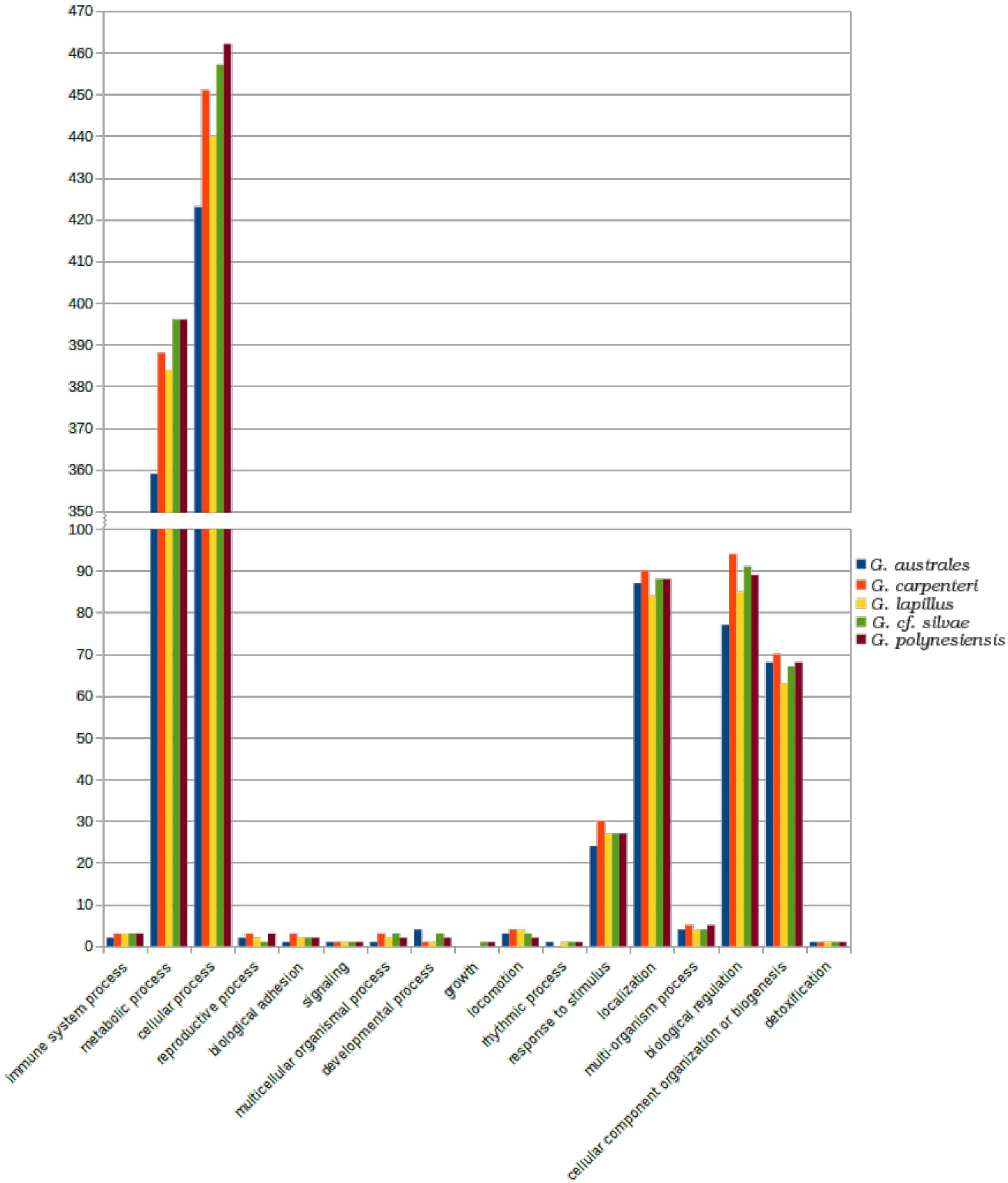

Figure 3: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 1 from Suppl. table 6.

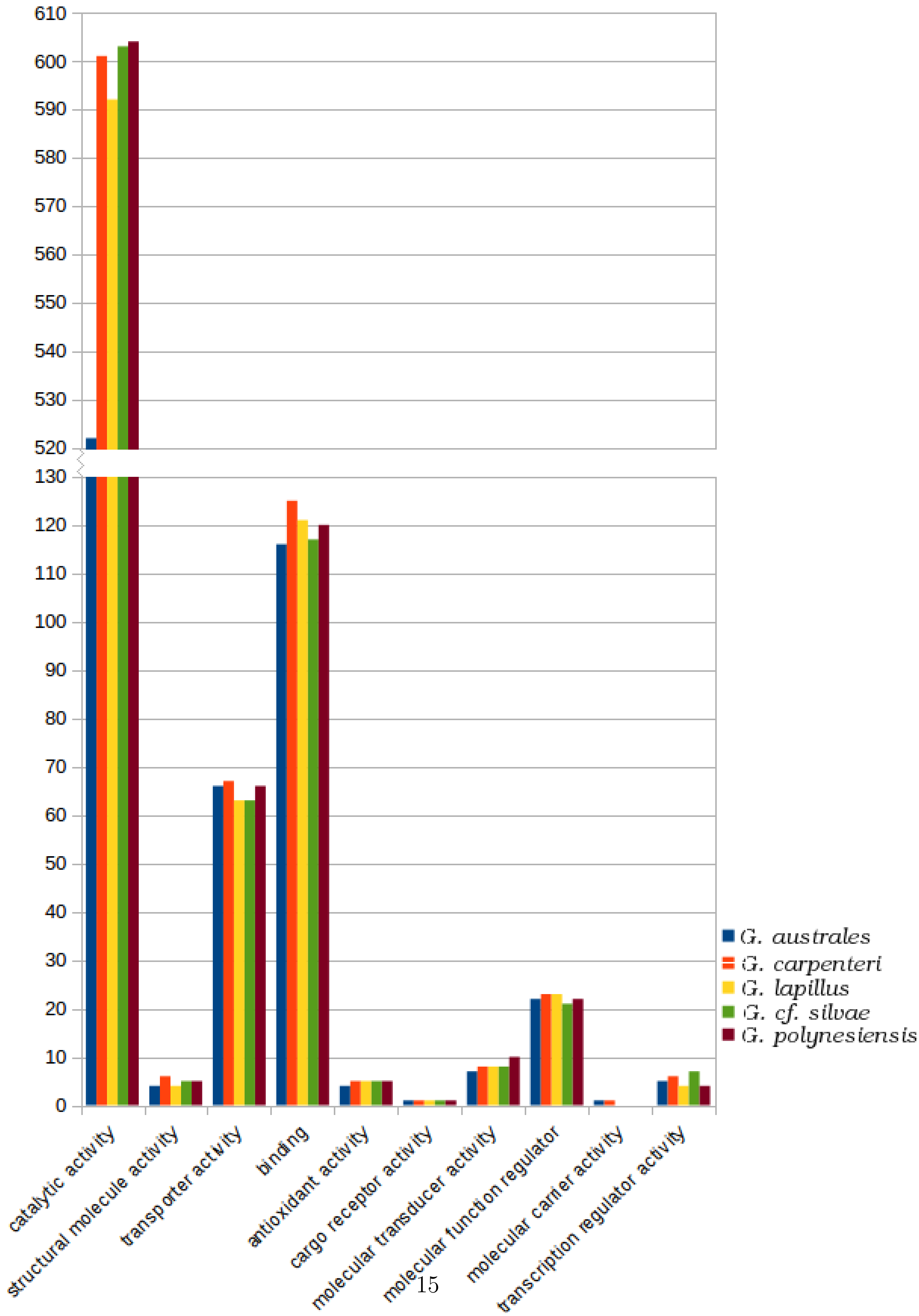

Figure 4: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 1 from Suppl. table 6.

At level 2 of GO annotations

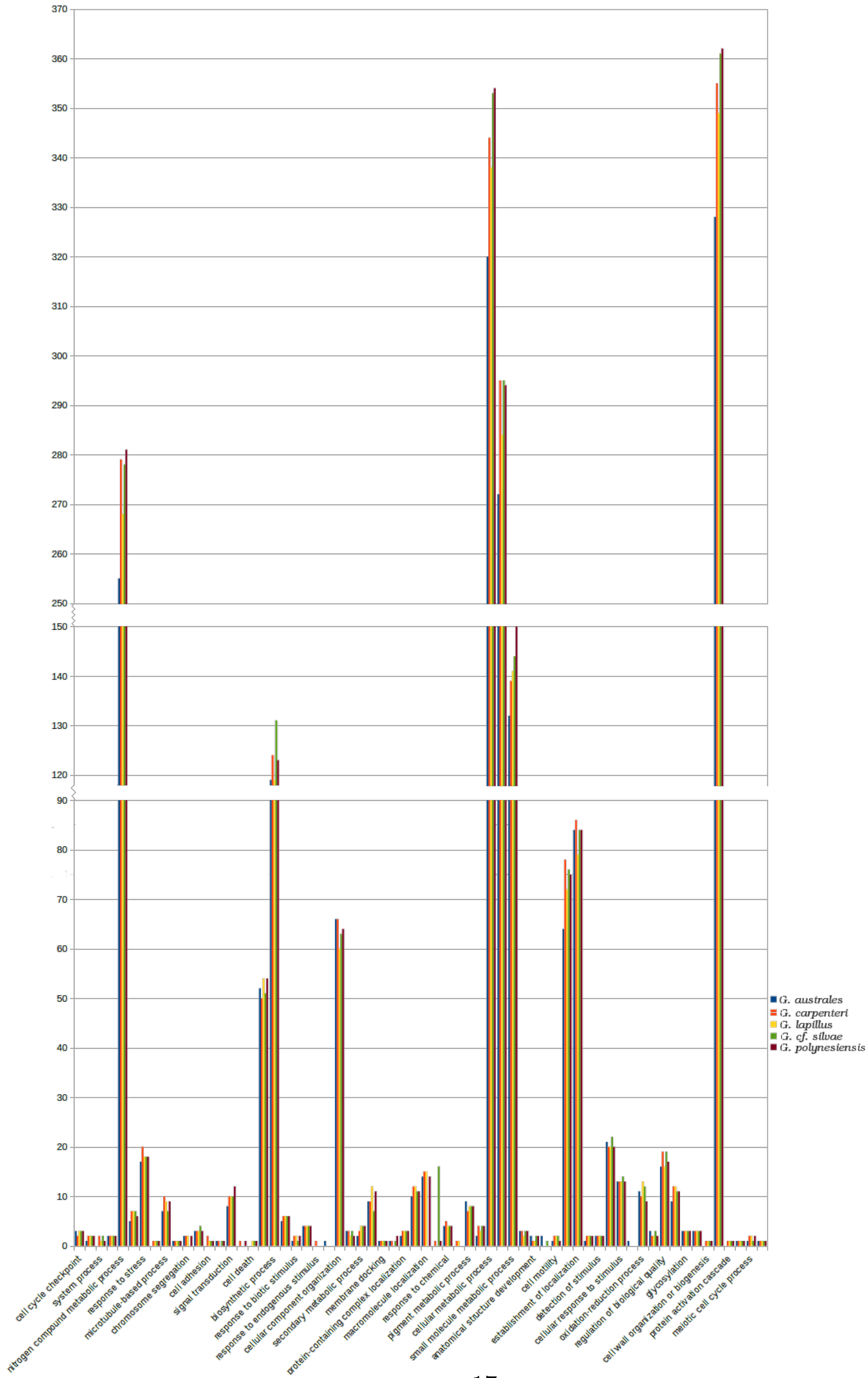

Figure 5: Summary of biological processes GO annotations between *Gambierdiscus* species at GOSUM level 2 from Suppl. table 7.

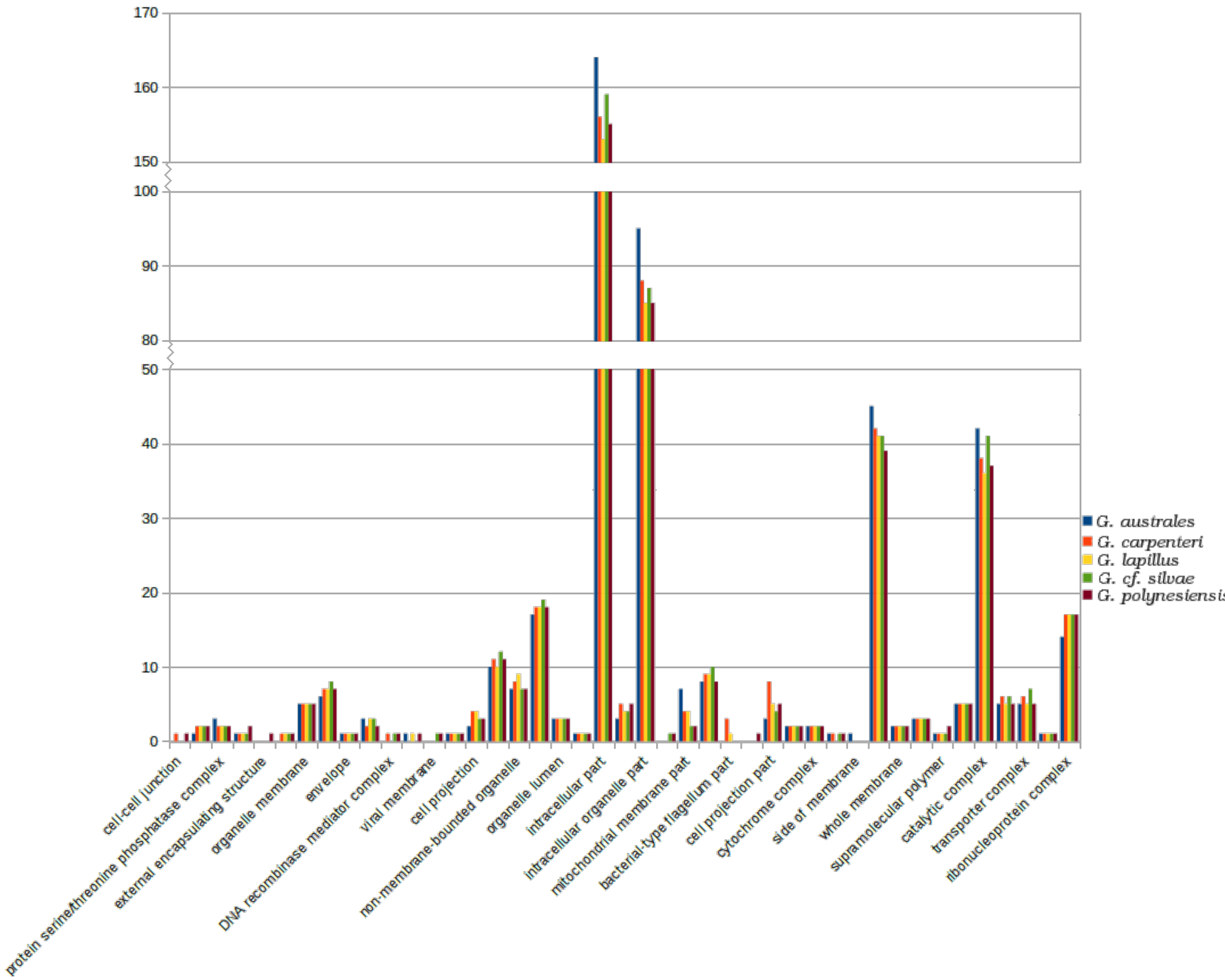

Figure 6: Summary of cellular GO annotations between *Gambierdiscus* species at GO-SUM level 2 from Suppl. table 7.

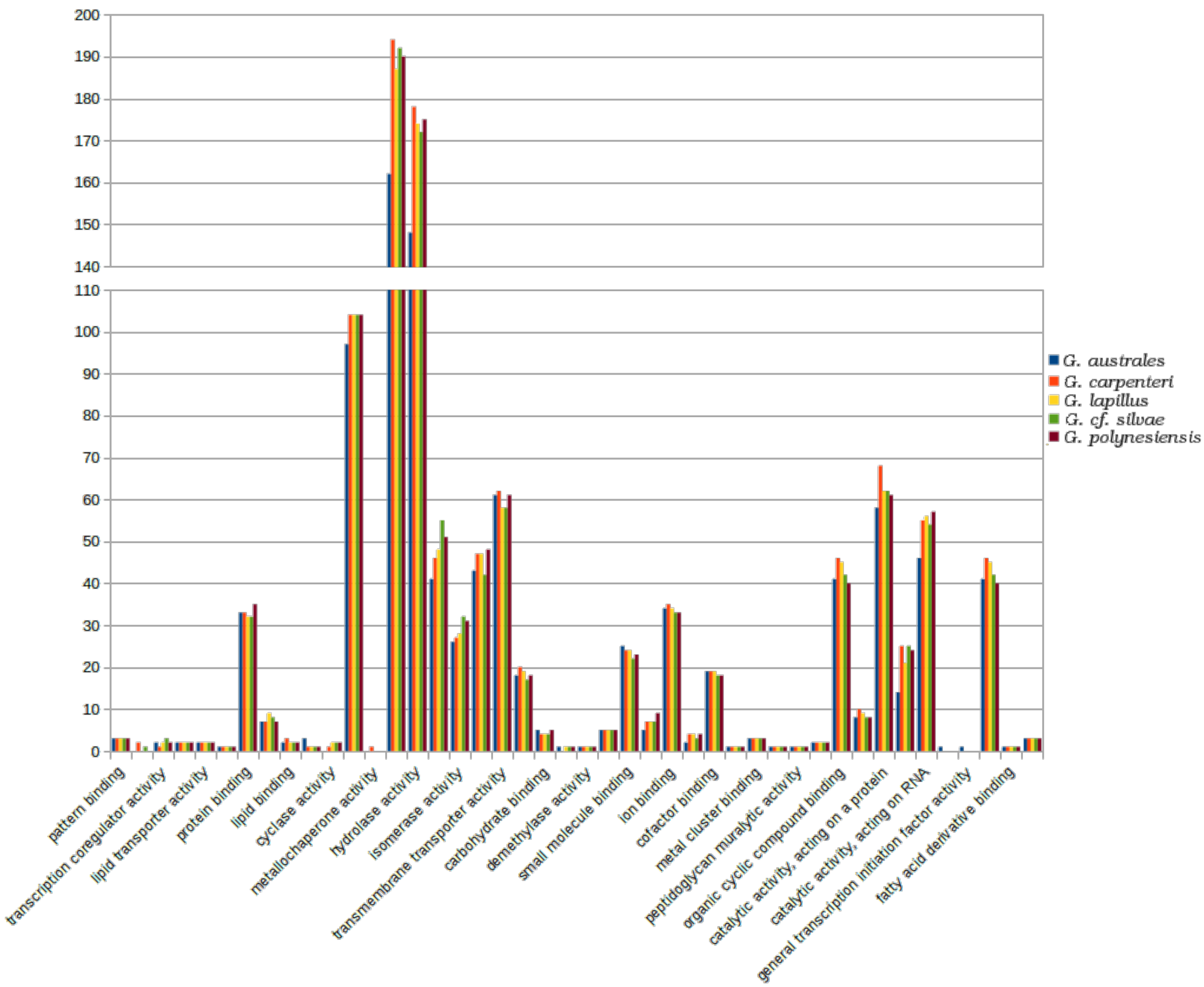

Figure 7: Summary of molecular GO annotations between *Gambierdiscus* species at GOSUM level 2 from Suppl. table 7.

4.2 Transcriptome similarity clustering

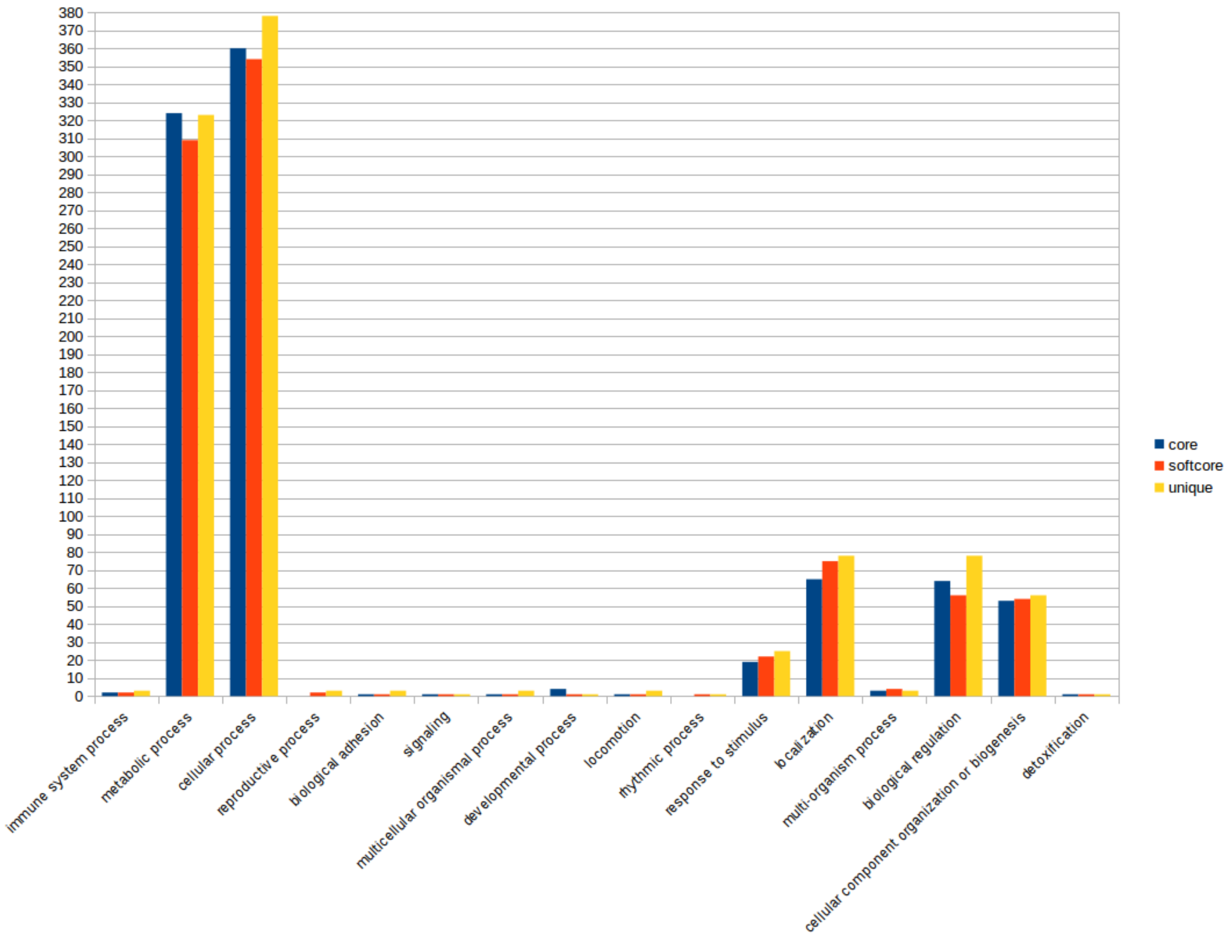

Figure 8: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 1.

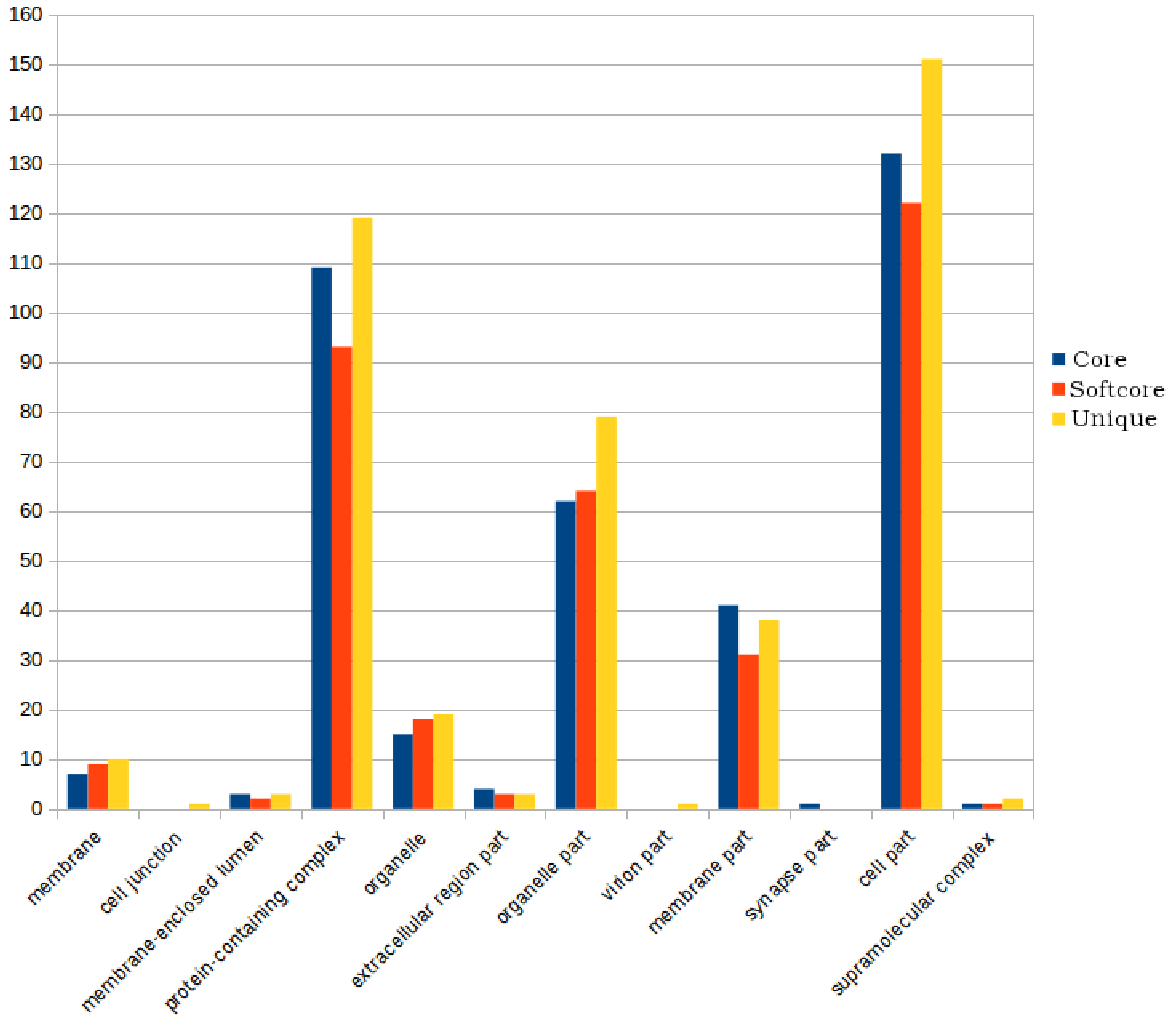

Figure 9: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 1.

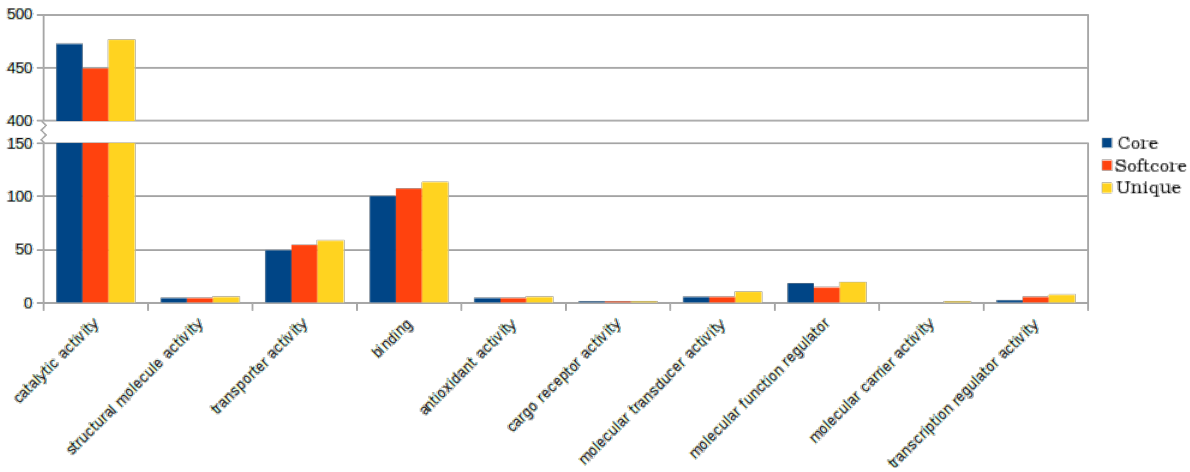

Figure 10: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 1.

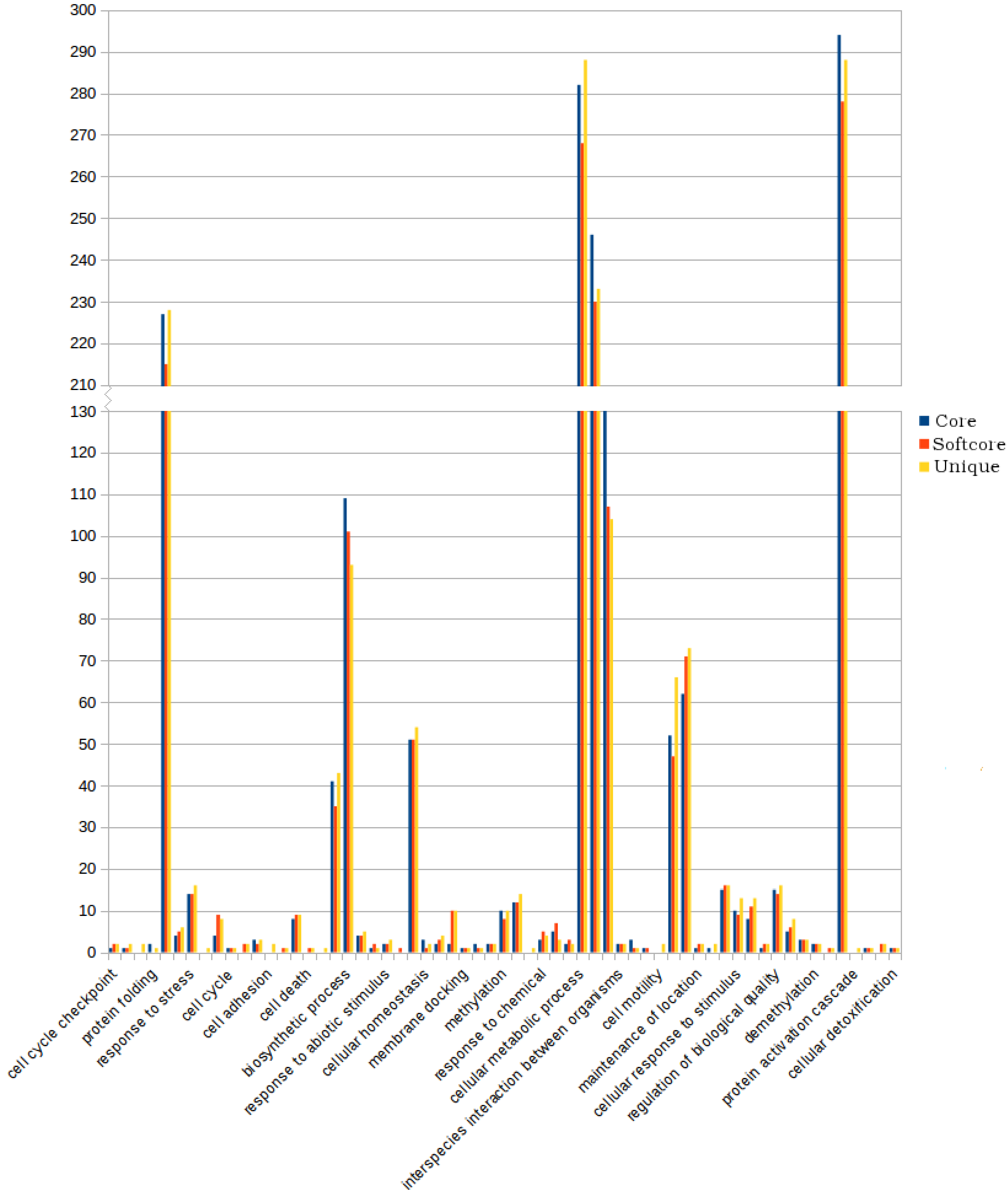

Figure 11: Summary of biological processes GO annotations between core, softcore and unique clusters at GOSUM level 2.

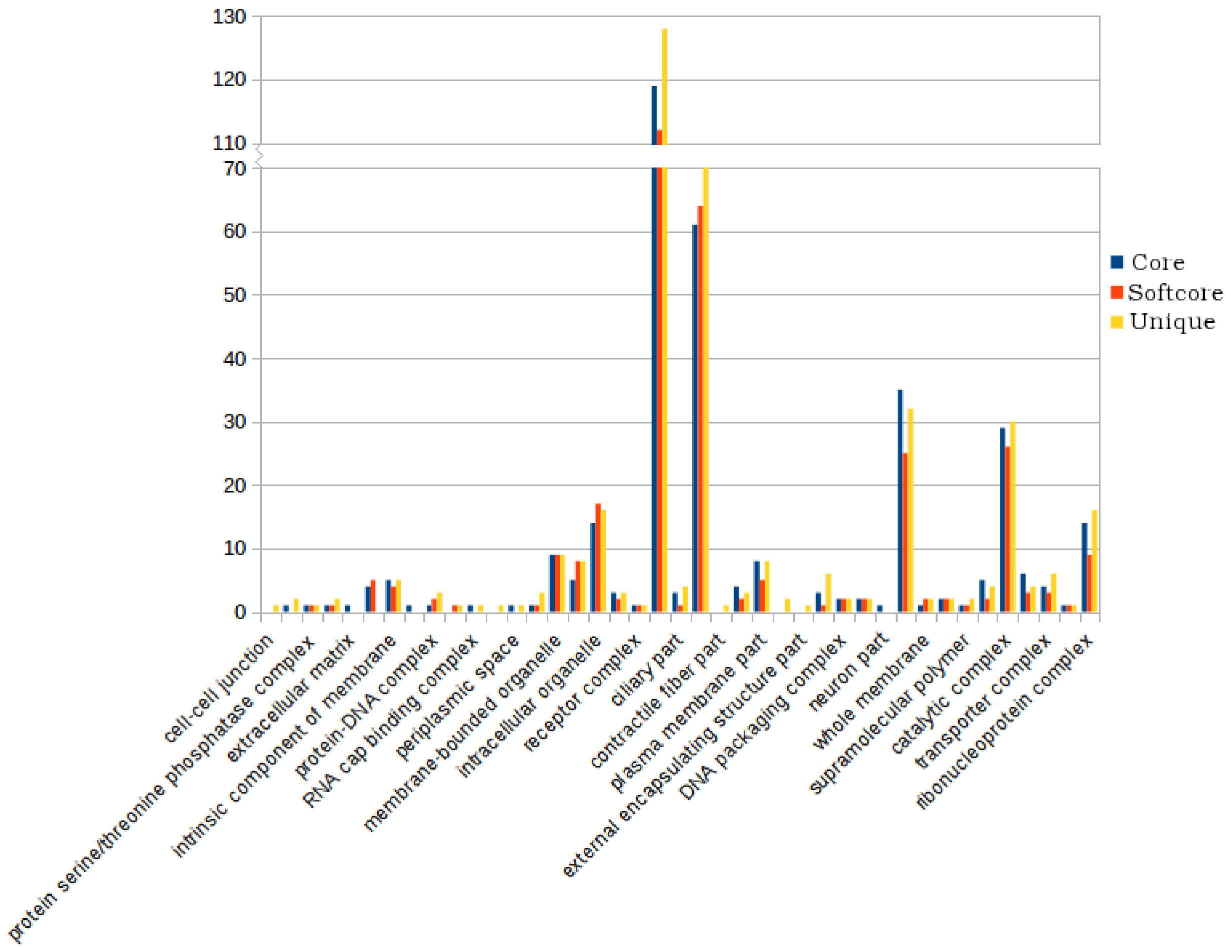

Figure 12: Summary of cellular GO annotations between core, softcore and unique clusters at GOSUM level 2.

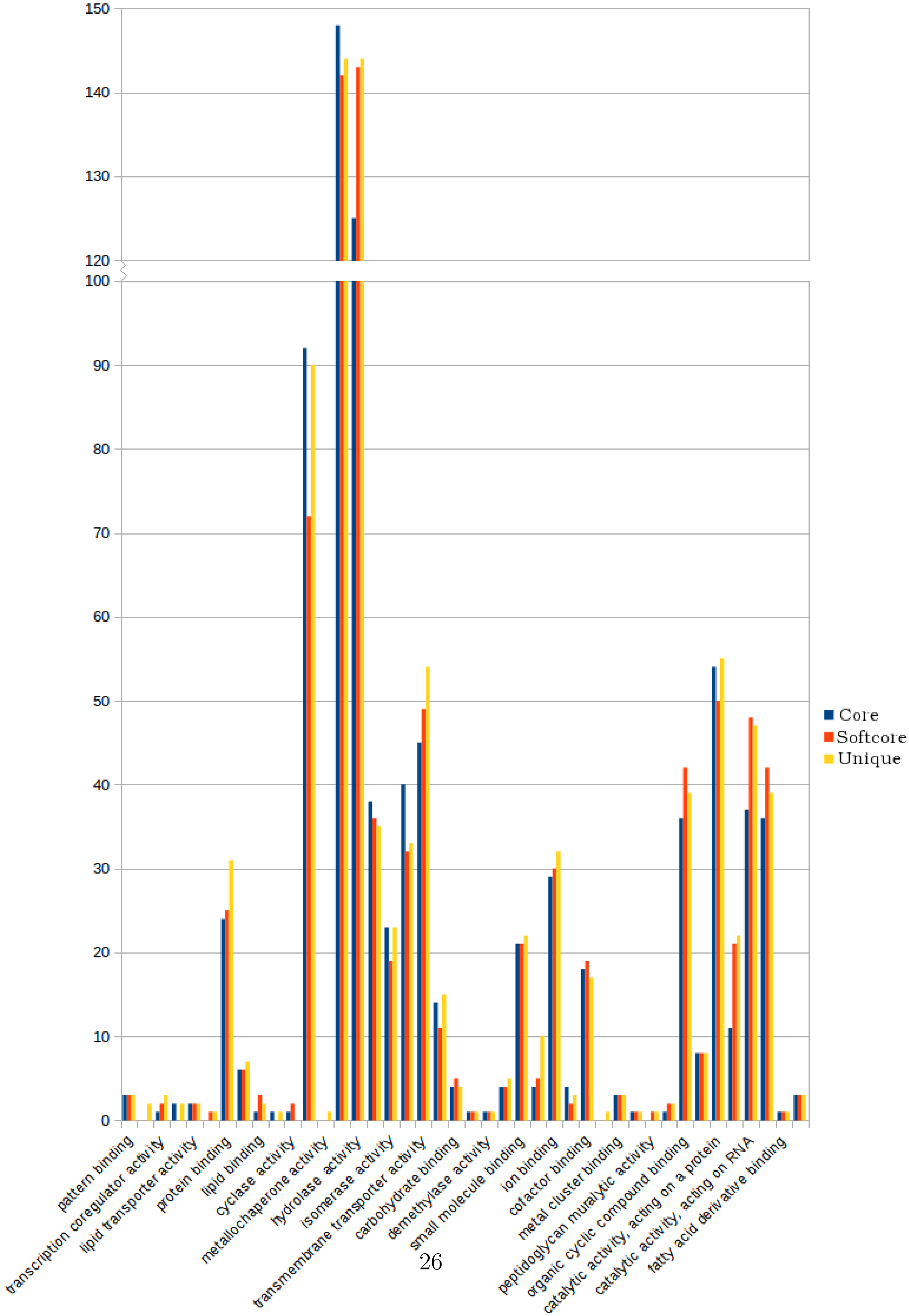

Figure 13: Summary of molecular GO annotations between core, softcore and unique clusters at GOSUM level 2.

4.2.1 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 8 & 9). 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.* cf. *silvae* 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?

4.2.2 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 8 & 9). The most prolific cluster in the softcore contained 163 contigs with unknown function, where *G. carpenteri* *G. lapillus*, *G. polynesiensis* and *G.* cf. *silvae* 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.

Table 3: LCA determination of clusters.

	Eukaryotic consensus	Eukaryotic unsure	Bacteria consensus	Bacteria unsure	Unknown between dbs	Unknown within db	Undetermined
Number of clusters	81,702	23,158	3,001	3,214	29,112	1,059	146,300
With dinoSL	341	76	11	12	81	6	759
with KS	0	8	0	5	255	0	7
with KS and dinoSL	0	0	0	0	0	0	0

Of the softcore, 14,035 clusters contained 10 or more contigs.

4.2.3 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.* cf. *silvae* 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.

4.3 Last common ancestor identification of contigs

Combined Swissprot and trEMBL

Table 4: basta trEMBL found in each *Gambierdiscus* transcriptome during processing.

Species	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*
Contigs	102,863	263,829	148,972	270,315	191,224
SwillProt					
SwissProt hits	62,240	176,000	109,662	171,741	129,913
BASTA positive ID	19,335	60,811	40,151	57,448	43,372
Eukaryotic origin	10,720	35,263	22,643	32,098	24,096
Bacterial origin	826	2,784	1,799	2,438	32,098
Unknown origin	7,709	22,429	15,471	22,571	17,072
trEMBL					
trEMBL hits	61,161	169,810	106,554	165,793	126,208
BASTA positive ID	37,067	106,960	71,100	103,053	106,960
Eukaryotic origin	25,015	65,986	44,320	62,274	49,516
Bacterial origin	654	2,213	1,404	2,101	1,688
Unknown origin	11,358	38,622	25,267	38,528	27,623
db differences					
contigs with LCA	37,294	108,160	71,768	104,252	79,692
db consensus	13,136	37,622	25,688	36,446	28,046
unknown plus LCA	5,821	21,399	13,434	19,247	14,158
LCA conflict, euk & bact	116	440	253	394	289

4.3.1 Unknown origin

To do:

- work out if PKS domains are within unknown
- may be bacterial origin - IF they have dinoSL, keep. If not, remove from core/pan analysis

4.3.2 Bacterial origin

To do:

- re-running with uniprot_trembl.fasta to see how percentage identity values differ to swissprot database
- merge trEMBL and swissprot databases and see how BASTA goes in comparison
- check if LCA is specific enough for Proteobacteria or gamma-Proteobacteria regarding Quorum sensing taxa
- make new directory with bacterial origin
- dinoSL search to see if any of bact origin are from dinos
- look if bact contigs found in unique or core clusters
- check if core bacteriome (how wanky is that word) or any species specific
- check for regional link of host association. Lapillus and silvae are from Heron Island from same collection trip, poly and australes are from Rarotonga collected 9 years apart, carp is from temperate Merimbula Merimbula)

4.4 Looking into toxin producers

not sure how valid an approach this following section is

Table 5: PKS active domains found in the *Gambierdiscus* species queries.

Active domain	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*	**Total contigs**	**# clusters**
ACP							
AT							
DR							
ET							
KS	130	195	150	221	154	850	314
KR							
TE							

4.4.1 Clusters that don't have *G. carpenteri* in

Rationale: This strain of carpenteri is the only one of the 5 which is a verified non-CTX producer, by LC-MS and bioassay.
To do:

- find clusters excluding carp
- look for clusters with higher number of contigs from poly and silvae as those are the two more toxic ones
- check for dinoSL and LCA of clusters

4.4.2 *G. polynesiensis* solo clusters

- number of clusters
- percentage annotated
- pathways present (another GOSUM adventure?)
- as *G. silvae* and to a much reduced extend, *G. lapillus*, also produce CTX, is the solo polynesiensis section relevant?

4.4.3 Polyketide synthase active domain search

KS domains. A total of 850 contigs were identified with KS domains which assembled into 314 clusters (table 5). 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.* cf. *silvae* with equally represented transcript numbers. Four contigs contained *G. polynesiensis* and *G.* cf. *silvae* only, one of which had a higher contig representation of *G. polynesiensis* than *G.* cf. *silvae*. *G. polynesiensis* was the only representative species in 71 clusters, of which three clusters contained 2 contigs and one cluster contained 3 contigs. *G.* cf. *silvae* 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.

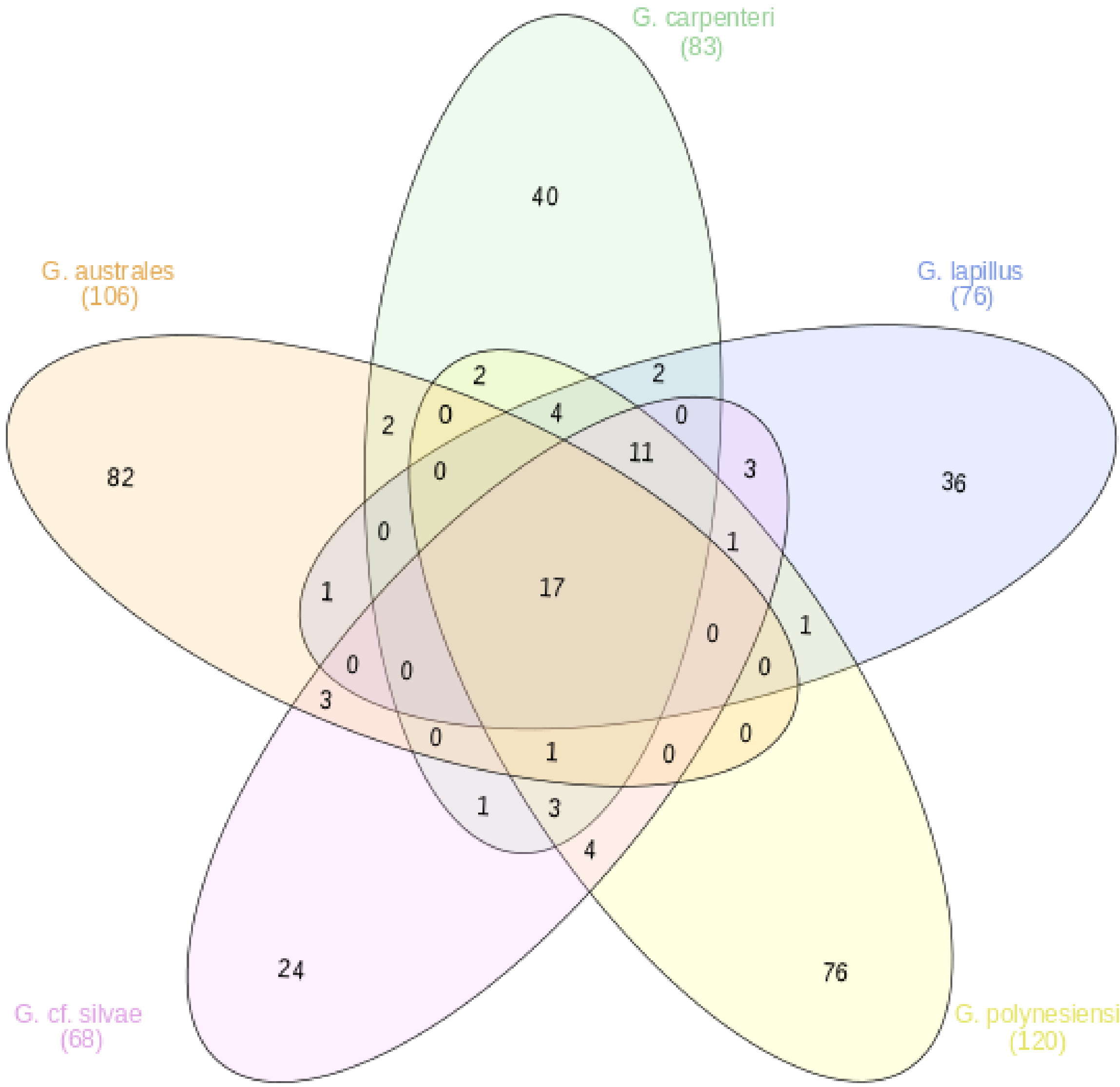

Figure 14: Venn diagram of species in KS clusters.

To do:

- are there any multi-domain transcripts?

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.

5.1 Core, soft-core and unique genes

The number of contigs and predicted peptides from this study was markedly variable between *G. australies*, 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.
Tim up to here, the rest is mostly based in the whole seq dataset and ideas on what to possibly look at
To go here:

- overall summary of study
- core and pan more likely to be accurate without being axenic unless same contamination **vs** removing bact LCA
- spliced leader sites really low. potentially interesting - the two highest ones are from same phylogenetic clade, while the other three are representatives from the other two main clades. Also poly and silvae are from separate seq runs, so not an artefact from that front.
- *G. australes* seq is quite bad in comparison as can be seen in the GOSUM figs and the comparative number of contigs, predicted proteins and softcore clusters

- **Tim** not sure if I can do something like Fig 2 - only 5 isolates to put in, and I think I need to look at that again with more sleep to work out what's going on and if I could transfer the concept., there are over 200,000 pan-tran clusters, I don't think I can work out whether they are genophyletic or monophyletic for that many

5.2 dinoSL

-differences between transcriptomes... either seq metod related, or taxa relate. I think silvae and poly had the most, which are from the same sub clade? - super low number of dinoSL found in libraries, not representative of all the transcripts [9] and they cite [2] as similar, but incompatible with findings by [44].. check zhang 2007 is it a detection thing, or genus/species specific differences

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

Bacterial association with host

- really depends what the basta results are and if anything interesting is found
- 'fundamental shift' in transcript expression observed in *A. tamarense* based on bacterial presence, much higher than N or P depletion [28]

discuss usefulness for future studies

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

discuss potential short comings

- from different seq runs and methods and seq depth may vary, especially *G. australes*
- intra-speces variation so one isolate per species may not be representative
- unknown if processes other than PKS play a role in toxin production

Conclusion

Supplementary

- need to add australes

Table 6: GO terms and number of contigs per species at GO ontology level 1.

GO acession	GO terms	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*
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 7: GO terms and number of contigs per species at GO ontology level 2, child terms of Table 6.

GO acession	GO terms	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*
		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 interaction 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 biological process	78	72	76	75
GO:0051234	establishment of localization	86	79	84	84
GO:0051235	maintenance of location	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 autophagic mechanism	2	2	3	2
GO:0065008	regulation of biological quality	19	16	19	17

GO:0065009	regulation of molecular 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 cascade	1	1	1	1
GO:0140029	exocytic process	1	1	1	1
GO:1903046	meiotic cell cycle process	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 serine/threonine phosphatase complex	2	2	2	2
GO:0019867	outer membrane	1	1	1	2
GO:0030312	external encapsulating 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 mediator 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 muralytic 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 8: GO terms and number of contigs found in core, softcore and pan-transcriptome of *Gambierdiscus* at GO ontology level 1.

GO acession	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 9: GO terms and number of contigs found in core, softcore and pan-transcriptome of *Gambierdiscus* at GO ontology level 2, childer to Table 8.

GO acession	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 serine/threonine phosphatase 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 mediator 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 membrane part	4	2	3
GO:0044459	plasma membrane part	8	5	8
GO:0044461	bacterial-type flagellum part	0	0	2
GO:0044462	external encapsulating structure part	0	0	1
GO:0044463	cell projection part	3	1	6
GO:0044815	DNA packaging complex	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 polymer	1	1	2
GO:0120114	Sm-like protein family complex	5	2	4

GO:1902494	catalytic complex	29	26	30
GO:1990204	oxidoreductase complex	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 muralytic 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 10: KS domains found per cluster and total number of contigs present.

Cluster ID	*G. australes*	*G. carpenteri*	*G. lapillus*	*G. polynesiensis*	*G.* cf. *silvae*	**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_UTSMEER9A3_Gambierdiscus-carpenteri_DN15967_c2_g1_i2.p1.faa	0	3	0	1	0	4
21082_MMETSP0766_Gambierdiscus-australes_DN32692_c0_g1_i1.p1.faa	0	1	0	2	1	4
195242_UTSMEER9A3_Gambierdiscus-carpenteri_DN17326_c2_g5_i1.p1.faa	0	1	1	1	1	4
83891_UTSMEER9A3_Gambierdiscus-carpenteri_DN13035_c1_g4_i1.p1.faa	[illegible]	1	1	1	1	4
99486_UTSMEER9A3_Gambierdiscus-carpenteri_DN13588_c0_g3_i1.p1.faa	[illegible]	1	1	1	1	4

328911_HG4_Gambierdiscus-lapillus_DN41464_c0_g1_i1.p1.faa	0	0	1	3	0	4
643864_HG5_Gambierdiscus-silvae_DN47931_c1_g3_i1.p2.faa	0	0	0	0	4	4
186957_UTSMER9A3_Gambierdiscus-carpenteri_DN16979_c3_g3_i1.p1.faa	0	1	1	1	0	3
193820_UTSMER9A3_Gambierdiscus-carpenteri_DN17268_c1_g8_i4.p1.faa	0	1	1	1	0	3
147284_UTSMER9A3_Gambierdiscus-carpenteri_DN15408_c1_g3_i2.p1.faa	0	1	1	1	0	3
116539_UTSMER9A3_Gambierdiscus-carpenteri_DN14227_c2_g1_i4.p1.faa	0	1	2	0	0	3
242595_UTSMER9A3_Gambierdiscus-carpenteri_DN9176_c0_g1_i3.p1.faa	0	1	2	0	0	3
524928_CG15_Gambierdiscus-polynesiensis_DN43543_c1_g1_i1.p1.faa	0	0	0	3	0	3
1040_MMETSP0766_Gambierdiscus-australes_DN11947_c0_g1_i1.p1.faa	1	0	0	0	2	3
38402_MMETSP0766_Gambierdiscus-australes_DN41494_c1_g1_i3.p1.faa	2	0	1	0	0	3
154624_UTSMER9A3_Gambierdiscus-carpenteri_DN15679_c0_g6_i1.p1.faa	0	2	0	0	0	2
63665_UTSMER9A3_Gambierdiscus-carpenteri_DN10182_c0_g1_i2.p1.faa	0	2	0	0	0	2
205876_UTSMER9A3_Gambierdiscus-carpenteri_DN17803_c0_g4_i1.p1.faa	0	2	0	0	0	2
224239_UTSMER9A3_Gambierdiscus-carpenteri_DN18618_c3_g6_i1.p1.faa	0	2	0	0	0	2
196786_UTSMER9A3_Gambierdiscus-carpenteri_DN17387_c2_g2_i1.p1.faa	0	1	0	1	0	2

131133_UTSMOER9A3_Gambierdiscus-carpenteri_DN14782_c2_g4_i3.p1.faa	0	1	0	0	1	2
19133_MMETSP0766_Gambierdiscus-australes_DN30780_c0_g2_i1.p1.faa	1	1	0	0	0	2
37007_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g7_i1.p1.faa	1	1	0	0	0	2
424979_CG150_Gambierdiscus-polynesiensis_DN34166_c0_g9_i1.p1.faa	0	0	0	2	0	2
358554_CG150_Gambierdiscus-polynesiensis_DN15070_c0_g1_i1.p2.faa	0	0	0	2	0	2
408901_CG150_Gambierdiscus-polynesiensis_DN32288_c2_g1_i1.p1.faa	0	0	0	2	0	2
479997_CG150_Gambierdiscus-polynesiensis_DN39607_c0_g2_i1.p1.faa	0	0	0	1	1	2
485470_CG150_Gambierdiscus-polynesiensis_DN40097_c0_g1_i2.p1.faa	0	0	0	1	1	2
258909_HG4_C_Gambierdiscus-lapillus_DN22432_c0_g1_i2.p1.faa	0	0	0	1	1	2
263811_HG4_C_Gambierdiscus-lapillus_DN25138_c0_g1_i1.p1.faa	0	0	1	0	1	2
319034_HG4_C_Gambierdiscus-lapillus_DN40675_c3_g1_i2.p1.faa	0	0	1	0	1	2
319505_HG4_C_Gambierdiscus-lapillus_DN40711_c1_g8_i1.p1.faa	0	0	1	0	1	2
1041_MMETSP0766_Gambierdiscus-australes_DN11947_c0_g2_i1.p1.faa	1	0	0	0	1	2
27066_MMETSP0766_Gambierdiscus-australes_DN36729_c0_g1_i1.p2.faa	1	0	0	0	1	2
274389_HG4_C_Gambierdiscus-lapillus_DN30113_c0_g1_i2.p1.faa	0	0	2	0	0	2

46553_MMETSP0766_Gambierdiscus-australes_DN42196_c9_g4_i1.p1.faa	2	0	0	0	0	2
148669_UTSMER9A3_Gambierdiscus-carpenteri_DN15462_c1_g7_i1.p1.faa	0	1	0	0	0	1
234513_UTSMER9A3_Gambierdiscus-carpenteri_DN23482_c0_g1_i1.p1.faa	0	1	0	0	0	1
63664_UTSMER9A3_Gambierdiscus-carpenteri_DN10182_c0_g1_i1.p1.faa	0	1	0	0	0	1
72166_UTSMER9A3_Gambierdiscus-carpenteri_DN1258_c0_g1_i1.p1.faa	0	1	0	0	0	1
210660_UTSMER9A3_Gambierdiscus-carpenteri_DN18011_c6_g4_i1.p1.faa	0	1	0	0	0	1
88291_UTSMER9A3_Gambierdiscus-carpenteri_DN13188_c2_g8_i2.p2.faa	0	1	0	0	0	1
235070_UTSMER9A3_Gambierdiscus-carpenteri_DN25711_c0_g1_i1.p2.faa	0	1	0	0	0	1
236919_UTSMER9A3_Gambierdiscus-carpenteri_DN33286_c0_g1_i1.p1.faa	0	1	0	0	0	1
234708_UTSMER9A3_Gambierdiscus-carpenteri_DN24051_c0_g1_i1.p1.faa	0	1	0	0	0	1
75892_UTSMER9A3_Gambierdiscus-carpenteri_DN12749_c1_g2_i3.p1.faa	0	1	0	0	0	1
207498_UTSMER9A3_Gambierdiscus-carpenteri_DN17871_c4_g9_i1.p1.faa	0	1	0	0	0	1
234298_UTSMER9A3_Gambierdiscus-carpenteri_DN22896_c0_g1_i1.p1.faa	0	1	0	0	0	1
84448_UTSMER9A3_Gambierdiscus-carpenteri_DN13053_c3_g3_i4.p1.faa	0	1	0	0	0	1
104611_UTSMER9A3_Gambierdiscus-carpenteri_DN13776_c4_g7_i1.p1.faa	0	1	0	0	0	1

242597_UTSMER9A3_Gambierdiscus-carpenteri_DN9176_c0_g2_i2.p2.faa	0	1	0	0	0	1
233698_UTSMER9A3_Gambierdiscus-carpenteri_DN2009_c0_g1_i1.p1.faa	0	1	0	0	0	1
115505_UTSMER9A3_Gambierdiscus-carpenteri_DN14189_c2_g12_i1.p1.faa	0	1	0	0	0	1
238946_UTSMER9A3_Gambierdiscus-carpenteri_DN4887_c0_g1_i1.p1.faa	0	1	0	0	0	1
208524_UTSMER9A3_Gambierdiscus-carpenteri_DN17914_c1_g3_i4.p1.faa	0	1	0	0	0	1
131131_UTSMER9A3_Gambierdiscus-carpenteri_DN14782_c2_g4_i1.p1.faa	0	1	0	0	0	1
215621_UTSMER9A3_Gambierdiscus-carpenteri_DN18221_c2_g6_i3.p1.faa	0	1	0	0	0	1
225926_UTSMER9A3_Gambierdiscus-carpenteri_DN18701_c1_g3_i2.p1.faa	0	1	0	0	0	1
239297_UTSMER9A3_Gambierdiscus-carpenteri_DN5390_c0_g1_i1.p1.faa	0	1	0	0	0	1
233616_UTSMER9A3_Gambierdiscus-carpenteri_DN19857_c0_g1_i1.p1.faa	0	1	0	0	0	1
208525_UTSMER9A3_Gambierdiscus-carpenteri_DN17914_c1_g3_i5.p2.faa	0	1	0	0	0	1
236171_UTSMER9A3_Gambierdiscus-carpenteri_DN30145_c0_g1_i1.p1.faa	0	1	0	0	0	1
241217_UTSMER9A3_Gambierdiscus-carpenteri_DN7872_c0_g1_i1.p1.faa	0	1	0	0	0	1
212813_UTSMER9A3_Gambierdiscus-carpenteri_DN18098_c3_g3_i2.p1.faa	0	1	0	0	0	1
147705_UTSMER9A3_Gambierdiscus-carpenteri_DN15422_c1_g3_i1.p1.faa	0	1	0	0	0	1

242594_UTSMER9A3_Gambierdiscus-carpenteri_DN9176_c0_g1_i2.p1.faa	0	1	0	0	0	1
86631_UTSMER9A3_Gambierdiscus-carpenteri_DN13131_c1_g1_i1.p1.faa	0	1	0	0	0	1
238247_UTSMER9A3_Gambierdiscus-carpenteri_DN38343_c0_g1_i1.p1.faa	0	1	0	0	0	1
212812_UTSMER9A3_Gambierdiscus-carpenteri_DN18098_c3_g3_i1.p1.faa	0	1	0	0	0	1
211703_UTSMER9A3_Gambierdiscus-carpenteri_DN18052_c3_g5_i1.p1.faa	0	1	0	0	0	1
239230_UTSMER9A3_Gambierdiscus-carpenteri_DN5288_c0_g1_i1.p1.faa	0	1	0	0	0	1
103957_UTSMER9A3_Gambierdiscus-carpenteri_DN13754_c3_g2_i4.p1.faa	0	1	0	0	0	1
462243_CG150Gambierdiscus-polynesiensis_DN37930_c0_g1_i2.p1.faa	0	0	0	1	0	1
355979_CG150Gambierdiscus-polynesiensis_DN10471_c0_g1_i1.p1.faa	0	0	0	1	0	1
524904_CG150Gambierdiscus-polynesiensis_DN43540_c1_g1_i2.p1.faa	0	0	0	1	0	1
471036_CG150Gambierdiscus-polynesiensis_DN38733_c0_g1_i1.p1.faa	0	0	0	1	0	1
527904_CG150Gambierdiscus-polynesiensis_DN43803_c0_g1_i1.p1.faa	0	0	0	1	0	1
494332_CG150Gambierdiscus-polynesiensis_DN40908_c1_g1_i1.p1.faa	0	0	0	1	0	1
475327_CG150Gambierdiscus-polynesiensis_DN39159_c1_g1_i1.p1.faa	0	0	0	1	0	1
446377_CG150Gambierdiscus-polynesiensis_DN36357_c3_g7_i1.p1.faa	0	0	0	1	0	1

415511_CG15_Gambierdiscus-polynesiensis_DN33112_c0_g1_i3.p1.faa	0	0	0	1	0	1
524930_CG15_Gambierdiscus-polynesiensis_DN43543_c1_g1_i4.p1.faa	0	0	0	1	0	1
500254_CG15_Gambierdiscus-polynesiensis_DN41444_c1_g3_i1.p1.faa	0	0	0	1	0	1
408903_CG15_Gambierdiscus-polynesiensis_DN32288_c3_g1_i1.p1.faa	0	0	0	1	0	1
211708_UTSMER9A3_Gambierdiscus-carpenteri_DN18052_c3_g5_i7.p1.faa	0	0	0	1	0	1
524905_CG15_Gambierdiscus-polynesiensis_DN43540_c1_g1_i3.p1.faa	0	0	0	1	0	1
528784_CG15_Gambierdiscus-polynesiensis_DN47453_c0_g1_i1.p3.faa	0	0	0	1	0	1
528223_CG15_Gambierdiscus-polynesiensis_DN44935_c0_g1_i1.p2.faa	0	0	0	1	0	1
362866_CG15_Gambierdiscus-polynesiensis_DN18821_c0_g1_i1.p1.faa	0	0	0	1	0	1
408898_CG15_Gambierdiscus-polynesiensis_DN32288_c1_g1_i1.p1.faa	0	0	0	1	0	1
473656_CG15_Gambierdiscus-polynesiensis_DN39000_c2_g2_i1.p1.faa	0	0	0	1	0	1
505619_CG15_Gambierdiscus-polynesiensis_DN41913_c1_g3_i1.p2.faa	0	0	0	1	0	1
357110_CG15_Gambierdiscus-polynesiensis_DN13123_c0_g1_i2.p2.faa	0	0	0	1	0	1
529123_CG15_Gambierdiscus-polynesiensis_DN48937_c0_g1_i1.p1.faa	0	0	0	1	0	1
419597_CG15_Gambierdiscus-polynesiensis_DN33575_c2_g1_i1.p1.faa	0	0	0	1	0	1

486622_CG150_Gambierdiscus-polynesiensis_DN40207_c2_g2_i2.p1.faa		0	0	1	0	1
518712_CG150_Gambierdiscus-polynesiensis_DN43045_c0_g2_i6.p1.faa		0	0	1	0	1
505617_CG150_Gambierdiscus-polynesiensis_DN41913_c1_g2_i1.p1.faa		0	0	1	0	1
419857_CG150_Gambierdiscus-polynesiensis_DN33604_c1_g1_i1.p1.faa		0	0	1	0	1
319033_HG4_Gambierdiscus-lapillus_DN40675_c3_g1_i1.p1.faa		0	0	1	0	1
505612_CG150_Gambierdiscus-polynesiensis_DN41913_c0_g1_i1.p1.faa		0	0	1	0	1
505621_CG150_Gambierdiscus-polynesiensis_DN41913_c1_g5_i1.p2.faa		0	0	1	0	1
368243_CG150_Gambierdiscus-polynesiensis_DN21805_c0_g1_i1.p1.faa		0	0	1	0	1
531066_CG150_Gambierdiscus-polynesiensis_DN7198_c0_g1_i1.p1.faa		0	0	1	0	1
411779_CG150_Gambierdiscus-polynesiensis_DN32643_c5_g2_i3.p2.faa		0	0	1	0	1
529709_CG150_Gambierdiscus-polynesiensis_DN51840_c0_g1_i1.p1.faa		0	0	1	0	1
424815_CG150_Gambierdiscus-polynesiensis_DN34144_c0_g1_i6.p1.faa		0	0	1	0	1
388829_CG150_Gambierdiscus-polynesiensis_DN29147_c0_g1_i1.p1.faa		0	0	1	0	1
528991_CG150_Gambierdiscus-polynesiensis_DN4849_c0_g1_i1.p2.faa		0	0	1	0	1
529886_CG150_Gambierdiscus-polynesiensis_DN52795_c0_g1_i1.p1.faa		0	0	1	0	1

517572_CG15_Gambierdiscus-polynesiensis_DN42942_c0_g1_i1.p1.faa	0	0	0	1	0	1
162319_UTSMER9A3_Gambierdiscus-carpenteri_DN15967_c2_g1_i1.p1.faa	0	0	0	1	0	1
486374_CG15_Gambierdiscus-polynesiensis_DN40177_c0_g2_i3.p1.faa	0	0	0	1	0	1
424977_CG15_Gambierdiscus-polynesiensis_DN34166_c0_g6_i1.p2.faa	0	0	0	1	0	1
480000_CG15_Gambierdiscus-polynesiensis_DN39607_c0_g2_i4.p1.faa	0	0	0	1	0	1
524933_CG15_Gambierdiscus-polynesiensis_DN43543_c1_g1_i7.p1.faa	0	0	0	1	0	1
529340_CG15_Gambierdiscus-polynesiensis_DN50363_c0_g1_i1.p1.faa	0	0	0	1	0	1
382787_CG15_Gambierdiscus-polynesiensis_DN27509_c0_g1_i1.p1.faa	0	0	0	1	0	1
455767_CG15_Gambierdiscus-polynesiensis_DN37290_c0_g4_i1.p1.faa	0	0	0	1	0	1
454667_CG15_Gambierdiscus-polynesiensis_DN37192_c1_g3_i1.p1.faa	0	0	0	1	0	1
505616_CG15_Gambierdiscus-polynesiensis_DN41913_c1_g1_i3.p1.faa	0	0	0	1	0	1
408904_CG15_Gambierdiscus-polynesiensis_DN32288_c3_g2_i1.p1.faa	0	0	0	1	0	1
519735_CG15_Gambierdiscus-polynesiensis_DN43127_c3_g5_i1.p1.faa	0	0	0	1	0	1
524932_CG15_Gambierdiscus-polynesiensis_DN43543_c1_g1_i6.p1.faa	0	0	0	1	0	1
419608_CG15_Gambierdiscus-polynesiensis_DN33575_c2_g2_i1.p1.faa	0	0	0	1	0	1

489214_CG150_Gambierdiscus-polynesiensis_DN40447_c0_g1_i2.p1.faa		0	0	1	0	1
407098_CG150_Gambierdiscus-polynesiensis_DN32057_c0_g1_i2.p1.faa		0	0	1	0	1
486620_CG150_Gambierdiscus-polynesiensis_DN40207_c2_g1_i2.p2.faa		0	0	1	0	1
529847_CG150_Gambierdiscus-polynesiensis_DN52688_c0_g1_i1.p1.faa		0	0	1	0	1
355910_CG150_Gambierdiscus-polynesiensis_DN1036_c0_g1_i1.p2.faa		0	0	1	0	1
419599_CG150_Gambierdiscus-polynesiensis_DN33575_c2_g1_i11.p1.faa		0	0	1	0	1
368244_CG150_Gambierdiscus-polynesiensis_DN21805_c0_g2_i1.p1.faa		0	0	1	0	1
528301_CG150_Gambierdiscus-polynesiensis_DN45312_c0_g1_i1.p1.faa		0	0	1	0	1
431157_CG150_Gambierdiscus-polynesiensis_DN34812_c2_g1_i1.p1.faa		0	0	1	0	1
429838_CG150_Gambierdiscus-polynesiensis_DN3467_c0_g1_i1.p1.faa		0	0	1	0	1
485799_CG150_Gambierdiscus-polynesiensis_DN40132_c0_g3_i1.p1.faa		0	0	1	0	1
449384_CG150_Gambierdiscus-polynesiensis_DN36673_c0_g1_i3.p1.faa		0	0	1	0	1
530384_CG150_Gambierdiscus-polynesiensis_DN55090_c0_g1_i1.p1.faa		0	0	1	0	1
357109_CG150_Gambierdiscus-polynesiensis_DN13123_c0_g1_i1.p2.faa		0	0	1	0	1
466543_CG150_Gambierdiscus-polynesiensis_DN38313_c1_g3_i1.p1.faa		0	0	1	0	1

367731_CG150_Gambierdiscus-polynesiensis_DN21547_c0_g1_i1.p1.faa		0	0	1	0	1
438506_CG150_Gambierdiscus-polynesiensis_DN35575_c0_g1_i7.p1.faa		0	0	1	0	1
491823_CG150_Gambierdiscus-polynesiensis_DN40690_c4_g5_i2.p1.faa		0	0	1	0	1
530249_CG150_Gambierdiscus-polynesiensis_DN54681_c0_g1_i1.p1.faa		0	0	1	0	1
661643_HG5_Gambierdiscus-silvae_DN57114_c0_g1_i1.p1.faa		0	0	0	1	1
601478_HG5_Gambierdiscus-silvae_DN43780_c7_g8_i1.p1.faa		0	0	0	1	1
567939_HG5_Gambierdiscus-silvae_DN35530_c0_g3_i1.p1.faa		0	0	0	1	1
593688_HG5_Gambierdiscus-silvae_DN42661_c0_g1_i1.p1.faa		0	0	0	1	1
540524_HG5_Gambierdiscus-silvae_DN20879_c0_g2_i1.p1.faa		0	0	0	1	1
649671_HG5_Gambierdiscus-silvae_DN48408_c0_g1_i4.p1.faa		0	0	0	1	1
620146_HG5_Gambierdiscus-silvae_DN45801_c1_g1_i1.p2.faa		0	0	0	1	1
589550_HG5_Gambierdiscus-silvae_DN41996_c3_g12_i1.p1.faa		0	0	0	1	1
643868_HG5_Gambierdiscus-silvae_DN47931_c1_g3_i5.p1.faa		0	0	0	1	1
657026_HG5_Gambierdiscus-silvae_DN48988_c0_g3_i1.p1.faa		0	0	0	1	1
589562_HG5_Gambierdiscus-silvae_DN41996_c3_g5_i1.p2.faa		0	0	0	1	1

608846_HG5_Gambierdiscus-silvae_DN44648_c2_g1_i1.p1.faa	0	0	0	0	1	1
593690_HG5_Gambierdiscus-silvae_DN42661_c0_g2_i3.p1.faa	0	0	0	0	1	1
550256_HG5_Gambierdiscus-silvae_DN27602_c0_g2_i1.p1.faa	0	0	0	0	1	1
608853_HG5_Gambierdiscus-silvae_DN44648_c2_g6_i1.p1.faa	0	0	0	0	1	1
559711_HG5_Gambierdiscus-silvae_DN32102_c0_g1_i2.p1.faa	0	0	0	0	1	1
575231_HG5_Gambierdiscus-silvae_DN38322_c1_g2_i1.p1.faa	0	0	0	0	1	1
591087_HG5_Gambierdiscus-silvae_DN42232_c1_g4_i1.p2.faa	0	0	0	0	1	1
657027_HG5_Gambierdiscus-silvae_DN48988_c0_g3_i2.p1.faa	0	0	0	0	1	1
540525_HG5_Gambierdiscus-silvae_DN20879_c0_g3_i1.p1.faa	0	0	0	0	1	1
601479_HG5_Gambierdiscus-silvae_DN43780_c7_g9_i1.p1.faa	0	0	0	0	1	1
589619_HG5_Gambierdiscus-silvae_DN42009_c0_g1_i3.p1.faa	0	0	0	0	1	1
596728_HG5_Gambierdiscus-silvae_DN43120_c1_g4_i4.p1.faa	0	0	0	0	1	1
254977_HG4_Gambierdiscus-lapillus_DN19871_c0_g1_i1.p1.faa	0	0	1	0	0	1
244474_HG4_Gambierdiscus-lapillus_DN10661_c0_g1_i1.p1.faa	0	0	1	0	0	1
354441_HG4_Gambierdiscus-lapillus_DN7536_c0_g1_i1.p2.faa	0	0	1	0	0	1

277633_HG4_Gambierdiscus-lapillus_DN31491_c0_g2_i1.p1.faa	C	0	1	0	0	1
312699_HG4_Gambierdiscus-lapillus_DN40082_c0_g1_i1.p1.faa	C	0	1	0	0	1
319501_HG4_Gambierdiscus-lapillus_DN40711_c1_g5_i1.p1.faa	C	0	1	0	0	1
244476_HG4_Gambierdiscus-lapillus_DN10661_c0_g2_i1.p1.faa	C	0	1	0	0	1
355588_HG4_Gambierdiscus-lapillus_DN9793_c0_g1_i1.p1.faa	C	0	1	0	0	1
351360_HG4_Gambierdiscus-lapillus_DN46619_c0_g1_i1.p2.faa	C	0	1	0	0	1
319490_HG4_Gambierdiscus-lapillus_DN40711_c1_g10_i1.p1.faa	C	0	1	0	0	1
249529_HG4_Gambierdiscus-lapillus_DN15767_c0_g4_i1.p1.faa	C	0	1	0	0	1
350445_HG4_Gambierdiscus-lapillus_DN4403_c0_g2_i1.p1.faa	C	0	1	0	0	1
249527_HG4_Gambierdiscus-lapillus_DN15767_c0_g2_i1.p1.faa	C	0	1	0	0	1
247959_HG4_Gambierdiscus-lapillus_DN14263_c0_g2_i1.p1.faa	C	0	1	0	0	1
354628_HG4_Gambierdiscus-lapillus_DN8017_c0_g2_i1.p1.faa	C	0	1	0	0	1
327310_HG4_Gambierdiscus-lapillus_DN41349_c0_g1_i2.p1.faa	C	0	1	0	0	1
245201_HG4_Gambierdiscus-lapillus_DN11411_c0_g1_i1.p1.faa	C	0	1	0	0	1
328839_HG4_Gambierdiscus-lapillus_DN41459_c1_g5_i1.p1.faa	C	0	1	0	0	1

332373_HG4_Gambierdiscus-lapillus_DN41718_c2_g1_i1.p1.faa	0	0	1	0	0	1
310068_HG4_Gambierdiscus-lapillus_DN39797_c2_g1_i1.p1.faa	0	0	1	0	0	1
355491_HG4_Gambierdiscus-lapillus_DN9601_c0_g2_i1.p1.faa	0	0	1	0	0	1
264742_HG4_Gambierdiscus-lapillus_DN25642_c0_g1_i1.p1.faa	0	0	1	0	0	1
310329_HG4_Gambierdiscus-lapillus_DN39821_c0_g4_i1.p1.faa	0	0	1	0	0	1
351275_HG4_Gambierdiscus-lapillus_DN46352_c0_g1_i1.p1.faa	0	0	1	0	0	1
298246_HG4_Gambierdiscus-lapillus_DN38038_c1_g5_i1.p1.faa	0	0	1	0	0	1
312700_HG4_Gambierdiscus-lapillus_DN40082_c0_g2_i1.p1.faa	0	0	1	0	0	1
245202_HG4_Gambierdiscus-lapillus_DN11411_c0_g1_i2.p1.faa	0	0	1	0	0	1
270811_HG4_Gambierdiscus-lapillus_DN28598_c0_g3_i1.p1.faa	0	0	1	0	0	1
311957_HG4_Gambierdiscus-lapillus_DN40004_c3_g1_i2.p1.faa	0	0	1	0	0	1
270397_HG4_Gambierdiscus-lapillus_DN2840_c0_g1_i2.p1.faa	0	0	1	0	0	1
351199_HG4_Gambierdiscus-lapillus_DN46156_c0_g1_i1.p1.faa	0	0	1	0	0	1
308197_HG4_Gambierdiscus-lapillus_DN39584_c0_g1_i1.p1.faa	0	0	1	0	0	1
354586_HG4_Gambierdiscus-lapillus_DN792_c0_g1_i1.p1.faa	0	0	1	0	0	1

350059_HG4_Gambierdiscus-lapillus_DN43172_c0_g1_i1.p1.faa	0	0	1	0	0	1
264744_HG4_Gambierdiscus-lapillus_DN25642_c0_g2_i1.p1.faa	0	0	1	0	0	1
27277_MMETSP0766_Gambierdiscus-australes_DN36847_c0_g3_i1.p1.faa	1	0	0	0	0	1
38401_MMETSP0766_Gambierdiscus-australes_DN41494_c1_g1_i2.p1.faa	1	0	0	0	0	1
14801_MMETSP0766_Gambierdiscus-australes_DN27057_c0_g3_i1.p1.faa	1	0	0	0	0	1
38397_MMETSP0766_Gambierdiscus-australes_DN41494_c0_g2_i1.p1.faa	1	0	0	0	0	1
19134_MMETSP0766_Gambierdiscus-australes_DN30780_c0_g3_i1.p1.faa	1	0	0	0	0	1
33030_MMETSP0766_Gambierdiscus-australes_DN39895_c0_g4_i1.p1.faa	1	0	0	0	0	1
35041_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g8_i1.p2.faa	1	0	0	0	0	1
40251_MMETSP0766_Gambierdiscus-australes_DN41766_c4_g24_i2.p1.faa	1	0	0	0	0	1
18687_MMETSP0766_Gambierdiscus-australes_DN30415_c0_g4_i1.p1.faa	1	0	0	0	0	1
18486_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g2_i1.p1.faa	1	0	0	0	0	1
61098_MMETSP0766_Gambierdiscus-australes_DN6084_c0_g1_i1.p1.faa	1	0	0	0	0	1
36998_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g1_i1.p1.faa	1	0	0	0	0	1
13967_MMETSP0766_Gambierdiscus-australes_DN26272_c0_g1_i1.p1.faa	1	0	0	0	0	1

19163_MMETSP0766_Gambierdiscus-australes_DN30800_c0_g7_i1.p2.faa	1	0	0	0	0	1
35033_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g1_i1.p1.faa	1	0	0	0	0	1
18548_MMETSP0766_Gambierdiscus-australes_DN30296_c0_g2_i1.p1.faa	1	0	0	0	0	1
36992_MMETSP0766_Gambierdiscus-australes_DN41205_c0_g1_i1.p1.faa	1	0	0	0	0	1
36641_MMETSP0766_Gambierdiscus-australes_DN41111_c0_g2_i2.p1.faa	1	0	0	0	0	1
37002_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g4_i1.p1.faa	1	0	0	0	0	1
28106_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g2_i1.p1.faa	1	0	0	0	0	1
72_MMETSP0766_Gambierdiscus-australes_DN10092_c0_g1_i1.p2.faa	1	0	0	0	0	1
15646_MMETSP0766_Gambierdiscus-australes_DN27800_c0_g3_i1.p1.faa	1	0	0	0	0	1
27067_MMETSP0766_Gambierdiscus-australes_DN36729_c0_g2_i1.p1.faa	1	0	0	0	0	1
24849_MMETSP0766_Gambierdiscus-australes_DN35413_c0_g3_i1.p1.faa	1	0	0	0	0	1
35035_MMETSP0766_Gambierdiscus-australes_DN40638_c0_g3_i1.p1.faa	1	0	0	0	0	1
35352_MMETSP0766_Gambierdiscus-australes_DN40756_c0_g1_i1.p1.faa	1	0	0	0	0	1
13100_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g3_i1.p2.faa	1	0	0	0	0	1
38396_MMETSP0766_Gambierdiscus-australes_DN41494_c0_g1_i1.p1.faa	1	0	0	0	0	1

27068_MMETSP0766_Gambierdiscus-australes_DN36729_c0_g3_i1.p1.faa	1	0	0	0	0	1
35359_MMETSP0766_Gambierdiscus-australes_DN40756_c1_g4_i1.p1.faa	1	0	0	0	0	1
30964_MMETSP0766_Gambierdiscus-australes_DN38922_c0_g1_i1.p1.faa	1	0	0	0	0	1
30406_MMETSP0766_Gambierdiscus-australes_DN38631_c0_g4_i1.p1.faa	1	0	0	0	0	1
36994_MMETSP0766_Gambierdiscus-australes_DN41205_c0_g3_i1.p1.faa	1	0	0	0	0	1
13103_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g6_i1.p1.faa	1	0	0	0	0	1
18485_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g1_i1.p1.faa	1	0	0	0	0	1
23610_MMETSP0766_Gambierdiscus-australes_DN34624_c0_g3_i1.p2.faa	1	0	0	0	0	1
18487_MMETSP0766_Gambierdiscus-australes_DN30257_c0_g2_i2.p1.faa	1	0	0	0	0	1
46548_MMETSP0766_Gambierdiscus-australes_DN42196_c9_g10_i1.p2.faa	1	0	0	0	0	1
15765_MMETSP0766_Gambierdiscus-australes_DN27921_c0_g3_i1.p2.faa	1	0	0	0	0	1
30418_MMETSP0766_Gambierdiscus-australes_DN38640_c0_g5_i1.p2.faa	1	0	0	0	0	1
17953_MMETSP0766_Gambierdiscus-australes_DN29796_c0_g1_i1.p1.faa	1	0	0	0	0	1
37006_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g6_i1.p2.faa	1	0	0	0	0	1
13101_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g4_i1.p1.faa	1	0	0	0	0	1

46559_MMETSP0766_Gambierdiscus-australes_DN42196_c9_g6_i3.p1.faa	1	0	0	0	0	1
17396_MMETSP0766_Gambierdiscus-australes_DN2924_c0_g1_i1.p1.faa	1	0	0	0	0	1
18547_MMETSP0766_Gambierdiscus-australes_DN30296_c0_g1_i1.p1.faa	1	0	0	0	0	1
7245_MMETSP0766_Gambierdiscus-australes_DN20901_c0_g2_i1.p1.faa	1	0	0	0	0	1
17216_MMETSP0766_Gambierdiscus-australes_DN29099_c0_g2_i1.p1.faa	1	0	0	0	0	1
30404_MMETSP0766_Gambierdiscus-australes_DN38631_c0_g1_i1.p1.faa	1	0	0	0	0	1
40250_MMETSP0766_Gambierdiscus-australes_DN41766_c4_g23_i1.p2.faa	1	0	0	0	0	1
41420_MMETSP0766_Gambierdiscus-australes_DN41862_c2_g4_i1.p1.faa	1	0	0	0	0	1
26875_MMETSP0766_Gambierdiscus-australes_DN36626_c0_g1_i1.p1.faa	1	0	0	0	0	1
28107_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g4_i1.p1.faa	1	0	0	0	0	1
14799_MMETSP0766_Gambierdiscus-australes_DN27057_c0_g1_i1.p1.faa	1	0	0	0	0	1
37863_MMETSP0766_Gambierdiscus-australes_DN41388_c0_g2_i1.p1.faa	1	0	0	0	0	1
38400_MMETSP0766_Gambierdiscus-australes_DN41494_c1_g1_i1.p1.faa	1	0	0	0	0	1
23609_MMETSP0766_Gambierdiscus-australes_DN34624_c0_g2_i1.p2.faa	1	0	0	0	0	1
10084_MMETSP0766_Gambierdiscus-australes_DN23190_c0_g1_i1.p1.faa	1	0	0	0	0	1

17955_MMETSP0766_Gambierdiscus-australes_DN29796_c0_g3_i1.p1.faa	1	0	0	0	0	1
13099_MMETSP0766_Gambierdiscus-australes_DN25567_c0_g2_i1.p2.faa	1	0	0	0	0	1
37003_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g4_i2.p1.faa	1	0	0	0	0	1
42718_MMETSP0766_Gambierdiscus-australes_DN41959_c8_g1_i1.p1.faa	1	0	0	0	0	1
52639_MMETSP0766_Gambierdiscus-australes_DN45380_c0_g1_i1.p1.faa	1	0	0	0	0	1
38398_MMETSP0766_Gambierdiscus-australes_DN41494_c0_g3_i1.p2.faa	1	0	0	0	0	1
28109_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g6_i1.p1.faa	1	0	0	0	0	1
24850_MMETSP0766_Gambierdiscus-australes_DN35413_c0_g4_i1.p1.faa	1	0	0	0	0	1
35358_MMETSP0766_Gambierdiscus-australes_DN40756_c1_g1_i1.p1.faa	1	0	0	0	0	1
15764_MMETSP0766_Gambierdiscus-australes_DN27921_c0_g2_i1.p1.faa	1	0	0	0	0	1
28108_MMETSP0766_Gambierdiscus-australes_DN37278_c0_g5_i1.p1.faa	1	0	0	0	0	1
18688_MMETSP0766_Gambierdiscus-australes_DN30415_c0_g5_i1.p1.faa	1	0	0	0	0	1
35357_MMETSP0766_Gambierdiscus-australes_DN40756_c0_g5_i1.p1.faa	1	0	0	0	0	1
35353_MMETSP0766_Gambierdiscus-australes_DN40756_c0_g1_i3.p1.faa	1	0	0	0	0	1
37009_MMETSP0766_Gambierdiscus-australes_DN41205_c1_g9_i1.p1.faa	1	0	0	0	0	1

15763_MMETSP0766_Gambierdiscus-australes_DN27921_c0_g1_i1.p1.faa	1	0	0	0	0	1
41417_MMETSP0766_Gambierdiscus-australes_DN41862_c2_g2_i1.p1.faa	1	0	0	0	0	1
28600_MMETSP0766_Gambierdiscus-australes_DN37530_c0_g2_i1.p1.faa	1	0	0	0	0	1
7400_MMETSP0766_Gambierdiscus-australes_DN21004_c0_g1_i1.p1.faa	1	0	0	0	0	1
35356_MMETSP0766_Gambierdiscus-australes_DN40756_c0_g4_i1.p1.faa	1	0	0	0	0	1
30416_MMETSP0766_Gambierdiscus-australes_DN38640_c0_g3_i1.p1.faa	1	0	0	0	0	1
18686_MMETSP0766_Gambierdiscus-australes_DN30415_c0_g2_i1.p1.faa	1	0	0	0	0	1
16842_MMETSP0766_Gambierdiscus-australes_DN28731_c0_g3_i1.p2.faa	1	0	0	0	0	1

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 polykeitde 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 *Emiliania* 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.