# Genetic study and geographical modelling distribution of the Ciguatera-Causing dinoflagellates, *Gambierdiscus and Fukuyoa* genera

**Àngels Tudó Casanova**
Master's degree in Biostatistics and Bioinformatics
Área del trabajo final

**Consultor/a:**
**Profesor/a: Paloma Pizarro Tobías**
**Carles Alcaraz**

Junio 2019

**Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)**

**A) Creative Commons:**

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento 3.0 España de Creative Commons

**B) GNU Free Documentation License (GNU FDL)**

Copyright © AÑO TU-NOMBRE.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free

**C) Copyright**

## FICHA DEL TRABAJO FINAL

| | |
|---|---|
| **Título del trabajo:** | *Descripción del trabajo* |
| **Nombre del autor:** | *Àngels Tudó Casanova* |
| **Nombre del consultor/a:** | *Paloma Pizarro Tobías* |
| **Nombre del PRA:** | Carles Alcaraz Cazorla |
| **Fecha de entrega (mm/aaaa):** | 06/2019 |
| **Titulación::** | *Máster Bioinformática y Bioestadística* |
| **Área del Trabajo Final:** | Microbiología, biotecnología y biología molecular |
| **Idioma del trabajo:** | *Inglés* |
| **Palabras clave** | *Microalge, ciguatoxin, phylogenetics, ciguatera* |


**Resumen**

*Gambierdiscus* y *Fukuyoa* son dos géneros de dinoflagelados que se encuentran principalmente en zonas tropicales, pero en las últimas décadas se han detectado en zonas temperadas o más frías. Parece ser que hay una expansión de estas microalgas mediada por el cambio climático.

Con este trabajo se quiere hacer una aproximación para examinar la diversidad genética de este género, ver si hay una relación genética y geográfica. Para ello se han utilizado herramientas clásicas de análisis genético. También se ha querido modelizar la presencia o ausencia de especies o de cada género, mediante modelos logísticos con un gran número de variables

Como resultado se han creado largo datasets de secuencias asociadas a coordenadas. Se ha podido ver la diversidad de ambos géneros y se ha podido calcular modelos logísticos para determinar una presencia o ausencia de las microalgas.Los trópicos albergan una gran diversidad de especies de estos dinolfagellados, pero podría haber índices de que se están expandiendo las especies. Por ahora con nuestros resultados, no se pueden concluir que haya una expansión, pero este trabajo es una primera aproximación para ver este tipo de expansiones de las microalgas. También hay un primer análisis con modelos logísticos basado en la presencia y ausencia de las microalgas para ver comprender qué variables determinan la distribución geográfica de las especies, estos análisis se pueden perfeccionar posteriormente con modelos más potentes.

## Abstract

*Gambierdiscus* and *Fukuyoa* are two genera of dinoflagellates found mainly in tropical zones, but in recent decades these species have been detected in cooler-temperate zones. It seems that there is an expansion of these microalgae mediated by climate change.

Aims of this work are study the genetic diversity of these genera, see if there is a genetic and geographical relationships and analyse the possible expansion. To this end, classical genetic analysis tools have been used. In addition, efforts have been done to model the presence or absence of species for each genera through logistic models with a large number of environmental variables.

As a result, long datasets of sequences associated with coordinates have been created. Throughout the created dataset has been analysed the diversity of both genera. Also, logistic models have been calculated to determine the presence or absence of microalgae. Tropical zones are hotspots of these dinoflagellates, but genetic indices of expansion might exist. For now, with our results, it is not possible to conclude that there is an expansion to cool areas, but this work is a first approach to observe this type of expansions in dinoflagellate. In addition, first analysis with logistic models based on the presence and absence of microalgae are studied, these analyses can be further completed in the future with more powerful models.

# INDEX

## List of figures



## List of tables



## Annexes

## 1.1 General description

Gambierdiscus and Fukuyoa genera are benthic dinoflagellates typically from the tropical and circumtropical areas [1]. Both genera live attached to macroalgae, sand and other substrates mainly in coastal areas. These two genera produce gambiertoxins (GTXs), which are the precursors of potent neurotoxins called ciguatoxins (CTXs)[2] . CTXs enter in the food web through herbivorous and they are bioaccumulated in the higher top-levels of the food web[2].

The consumption of seafood contaminated with CTXs, may cause a disease called Ciguatera Food Poisoning (CFP), which is one of the most seafood-borne illnesses associated with fish consumption in worldwide, it is estimated to affect more than 25,000-500,000 persons per year [3].However, It is estimated that only 10%-20% of CFP cases are reported. The symptoms of intoxication are typically gastrointestinal, cardiovascular and neurological disturbances, which can last days, weeks and months[4]. Fatal cases are rare but, they have been described[5].In communities from tropical areas where diet is based on fish, CFP has been an important influence over fishing practices, dietary practices and migration patterns[3]. Economic impact is noteworthy, but a worldwide estimation does not exist. Although, in the United States, it was estimated at US$21 million annually for the period from 1987 to 1992 [3].

In recent years, CFP cases are increasing and expanding to non-endemic areas[6] probably mediated by climate change. In Europe, since 2004 CFP outbreaks appeared in Macaronesia (Canary and Madeira archipelago). After several poisonings, European Food Safety Authority (EFSA) declared CFP as an emergent disease in Europe and a priority issue to study. In a short time, authorities have financed studies on epidemiology of ciguatera, improving detection methods of CTXs in seafood and microalgae, reporting populations of CTXs producers. Identification of species by light microscopy and scanning electron microscopy (SEM) is very difficult, therefore identification is based on molecular biology.

Last years, new methodologies, records and revisions have caused of a constant

updating of taxonomy of Gambierdiscus and Fukuyoa genera[7–9].

Research shows not all species produce the same toxins. In addition, some strains present more toxic compounds than the others, and some them seem to be non-producers[10]

It is suggested that toxic production (fg CTX3C equiv. $cell^{-1} \cdot d^{-1}$) depends on genetically more than environmental parameters [11].Therefore, well identification of species is necessary in order to evaluate the local risk of Ciguatera.

Moreover, identify those areas that might be potencial locations for high toxic species are also crucial to evaluate the future risk of Ciguatera. By modelling presence and absence of species is possible to know information about the geographical distribution and predict events for species, for instance, to predict invasion and proliferation under climate changes scenarios[12]. In algal research, modelling is focus mainly on proliferations, to predict the abundance of determinate species under influence of environmental conditions. Good models for geographical distribution of the most toxic species could be crucial to reduce Ciguatera risk in local zones.

The present study contains two parts clearly differentiates. The first part is focused on the genetic diversity of the *Gambierdiscus* and *Fukuyoa* genera, phylogenetics analysis of strains froem databases and strains from Institut de Recerca i Tecnologies Agroalimentàries (IRTA) were performed. The diversity of the current species was analysed to understand the genetic diversity of both genera in worldwide. Phylogenetics studies are conducted with nuclear-encoded ribosomal RNA gene (rDNA) (LSU D8-D10, LSU D1-D3 and ITS1-5.8 ITS2). We particularly placed emphasis with *Gambierdiscus australes* from Europe, which is a common species in the Balearic islands (western Mediterranean Sea) and the Canary islands (North Atlantic Ocean) and is only species reported in the western Mediterranean Sea.

The second part of the study is focused on characterize the environmental conditions where species of CTX-producers are distributed. In addition, a model logistic based on presence and absence was performed. Revision of the literature were done to compile all locations, where CTX-producers were reported. Model is based on presence and

pseudoabsence. Pseudoabsence is an artificial data, that represents species which were not present in the sampling point [13]. To confirm absences in marine species is very difficult, particularly depth of the samplings could represent a bias. Historically, samplings of *Gambierdiscus* and *Fukuyoa* genera are proceeded by apnoea a few meters of depth and findings in deeper zones have been reported by chance. Last decades, modelling with presence and absence data combined with environmental data using geographical information system (GIS) technology are increased and it is possible to model with different approximations, since simple such as Randon Forest (RF) to more complex such as Maxent Models (maximum entropy modeling) [14,15]. In this study has been used a multiple logistic regression, involving a logit link and binomial error distribution [16,17].

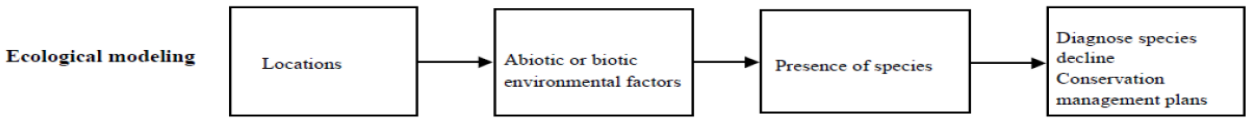



**Figure 1. Ecological modelling from Manel et al. 2001[18].**

This study is a first genetic study with large dataset of CTX-producers (n=434), with strains from databases and new sequences from Europe. Moreover, it is the first study in order to model the geographical distribution of *Gambierdiscus* and *Fukuyoa* genera and large dataset of variables (n=311) are analysed.

## 1.2 Objetives

1. Analyze the genetic diversity of *Gambierdiscus* and *Fukuyoa* genera in the world.
2. Understand and study the possible expansion of the *Gambierdiscus australes* in Europe.
3. Analyze the possible relationships between the Canary Islands and Balearic islands populations of *G.australes*.
4. Identify important environmental variables for the presence of *Gambierdiscus* and *Fukuyoa* genera.
5. Model species distribution in areas where CTX-producers are present.

## 1.3 Approaches and strategies.

Study combines information of strains from the *Gambierdiscus* and *Fukuyoa* genera from databases and literature and the work of the collection of microalgae from IRTA. Laboratory work have been done in parallel with this study and database was constantly updated. Analyses has been performed by classical programs although, is also performed by R software [19].

Particularly, the second part of this work have been proceeded after studying possible models for modelling the presence or absence of species; we decided to use logistic regression, which is a model used in geographical studies widely and which is feasible to work in 4 months.

## 1.4 Planning of the project

As It is mentioned above, some sequences have been obtained from databases, and others from IRTA. IRTA sequences have been sequenced in parallel with the data analysis. This methodology has been a handicap because all time database was updating in order to have as much information as possible. For these reasons loads of analysis have been doing to have new information.

Some results of analyses data have provoked that some goals which were planned previously are dismissed. For example, wide genetic analysis with region ITS rDNA was discarded due to lack of sequences in the databases.

First was planned to work only with *G. australes* species, but after PEC1 with information from data sets and calibration of time, previous ambitious goals were reduced and dataset was amplify to all species and to work basically for phylogenetic analysis with only one molecular marker. Planning is showed with next Gantt chart (fig.2)

**Figure 2. Chart of Gantt of the project**

## 1.5 Goals achieved

From the study a database with geographical information and environmental information has been created. Moreover, an updated global analysis of the genetic diversity and distribution of CTX-producers species have been achieved.

Finally, good logistic models to predict presence and absence of some species have been obtained.

## 1.6 Brief description of the other chapters of the memory

This study is divided in two parts clearly well defined. The first part constitutes genetic analysis of ciguatoxins-producer species (*Gambierdiscus* spp. and *Fukuyoa* spp.). Data from the literature of worldwide and new data of Europe from IRTA have been combined. Specifically in the analysis has taken in account *G. australes,* which is common species in the Western Mediterranean Sea and Northerm Athlantic Sea.

After genetic diversity analysis, from all sampling points where ciguatoxins-producers have been found logistic models have been estimated when it was possible.

# Chapter 1

# 2. Methods

## 2.1 Creation of dataset for phylogenetics analysis

### 2.2.1 Genetic data

Sequences from LSU rDNA D1-D3 region, LSU rDNA D8-D10 and ITS1-5.8 rDNA-ITS2 of all species of *Gambierdiscus* and *Fukuyoa* genera were obtained from GenBank database National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov and from IRTA. Selection of sequences for the analysis was revised for each article and verified when was needed by blast (Basic Local Alignment Search Tool). Two approaches of downloading sequences was performed, directly from databases and using *"*rentrez*"* R package (see in annex 1). In addition, all relevant information about sequences such as organism, amplified region, origin, sampling point, coordinates, article, and authors was compiled. Sequences from IRTA were cleaned and edited by Bioedit v.7.05[20].

### 2.2.2 Obtaining coordinates from sampling points

After the creation of datasets with sequences of *Gambierdiscus* and *Fukuyoa* genera from worldwide, coordinates from the sampling point of each strain was collected were compile. Coordinates were taken directly from the articles or by inferring from the description of the area in the articles. Coordinates were added to the dataset with information of strains from previous work. Data contained number accession, name of species and coordinates. All coordinates were uploaded in Google earth Pro (v.7.3.2.5776) in order to verify manually the location of each strain. As a result, 264 strains were located with coordinates (latitude and longitude) from different parts of the world.

## 2.2 Genetic data analysis

After creation of the datasets, genetic diversity analyses were performed with sequences of LSU rDNA D1-D3 region, LSU rDNA D8-D10 and ITS1-5.8 rDNA-ITS2 of all species of *Gambierdiscus* and *Fukuyoa*. For that sequences were aligned through

*Clustal* W[21] and "ape" R package. *Clustal W* is a free and intuitive software to align the sequences, which, it is possible to work with multiple sequences. Also, *Mafft*, *muscle and online* program was considered to be used. Sequences were edited with "ape" and "seqir R" packages (see in annex 2).

To sum up, final datasets were:

- D8-D10 with 434 sequences of *Gambierdiscus* spp. and *Fukuyoa* spp. (623 pb).
- D8-D10 with 100 sequences of *G. australes* (748 pb).
- D1-D3 with 47 sequences of *G. australes* (750 pb).
- ITS1-5.8 rDNA-ITS2 with 42 sequences of *Gambierdiscus* spp. and *Fukuyoa* spp. from worldwide (490 pb).

### 2.2.1 Estimation of the best evolution model.

For the phylogenetic analysis, firstly, the most appropriate model of evolution was determined by two approximations; through ModelTest () package "phangorn" with Akaik Information Criterion (AIC) and Bayesian Information Criterion (BIC). Model was studied for all of sequences of LSU markers and another for *G. australes* dataset.

### 2.2.2 Genetic distances (annex 3)

Genetic distances for the dataset of molecular markers D8-D10 and D1-D3 rDNA were estimated using uncorrected genetic distance (UGD) using "ape" and "phangorn" R packages and with software MEGA7. R packages and MEGA7 do not admit complex models therefore was not possible to calculate distances with the model GTR+G. Genetic distance for each taxon was save in excel files. In addition to visualize distances, distance trees were performed with "ape", "phangorn" and MEGA7.

### 2.2.3 Estimation of geographical distances (annex 2 (section 2.2)

Geographical distances measured in (Km) were obtained from coordinates (latitude and longitude) were estimated with "geosphere" from R packages. Then geographical distances were saved in excel files.

### 2.2.3 Correlation between genetic distance and geographical distances (annex 2 section 2.2)

Correlations were proceeded by mantel test of “vegan” R package only for *G.australes* populations.

### 2.2.4 Genetic diversity (annex 2, section 2.1)

As a consequence of the long time to obtain a complete data set for analysis *Gambierdiscus* and *Fukuyoa* genera for molecular markers, only dataset of D8-D10 rDNA region was considered in this part. Finally, dataset contains 434 sequences of length of 592 pb. Genetic diversity for each specie and for each genus was analysed with DNAsp[22] for D8-D10 LSU rDNA region considering that this region has more sequences than the other molecular markers. Previously, a data set with no ambiguities and gaps, was created. The final length of studied sequences changes for each species or genus. Results from DNAsp were summarized with a table (see in results table1).

Parameters were studied:

- Number of polymorphic segregation sites (S.pol)
- Nucleotide diversity pi (π)
- Number of haplotypes (n° H)
- Haplotype gene diversity (H)
- Fu F’s D statistic (Fu and Li 1993) [23]
- Fu F’s statistic (Fu and Li 1993) [23]

## 2.3.Phylogenetic analysis (annex2)

After aligning the sequences and the selection of evolution model was chosen, phylogenetic analyses were proceeded. Trees based on genetic distance estimated by methods: neighbour joining method (NJ) and UPGMA (unweighted pair group method with arithmetic mean). In additon, phylogenetic trees were obtained by Maximum Likelihood (ML) and Bayesian Inference (BI). For phylogenetic analysis of all species of *Gambierdiscus* and *Fukuyoa*, a dinoflagellate *Coolia monotis* was used as an outgrup. Specifically, for phylogenetic analysis with strains of *G.australes* the outgrup *F. paulensis*.

I) Phylogenetic trees based on Distance Methods

Genetic distance pairwise genetic distance was estimated with two approximations, “ape” and “phangorn” packages of R and MEGA7. Phylogenetic trees were obtained with neighbour joining method (NJ) and UPGMA (unweighted pair group method with arithmetic mean). Although as a result of the estimation of evolution model was a complex model such as Generalised time-reversible model (GTR+G), It is not possible to calculate distances with this type of models. Therefore, genetic distances were estimated with K80 model, which assumes that nucleotides mutate with the same probability. Some large trees were edited with iTol (interactive tree of life) https://itol.embl.de/tree to make them easy interpretation.

II) Phylogenetic trees based on Maximum likelihood

Trees based on Maximum Likelihood (ML) were obtained with MEGA7. Parameters were evolution model GTR+G and the option *complete delection*. This option eliminates all positions with gaps in the sequence, being the most conservative option.

# 3. Results

## 3.0. Creation of dataset

For each strain information of Genbak code, isolate code, species, origin, publication and authors of publication was compiled

Summary of dataset:

D8-D10 with 434 sequences of *Gambierdiscus* spp. and *Fukuyoa* spp. (623 pb).

D8-D10 with 100 sequences of *G. australes* (748 pb).

D1-D3 with 47 sequences of *G. australes* (750 pb).

ITS1-5.8 rDNA-ITS2 with 42 sequences of Gambierdiscus spp. and Fukuyoa spp. from worldwide (490 pb).

To work with dataset of ITS1-5.8 rDNA-ITS2 was dismissed because dataset contained few strains for both genera, hence analysis will be with other molecular markers. For

both genera was used LSU D8-10 region and for *G.australes* have been used LSU D8-10 and D1-D3 regions. At least sequences and information of 17 species and their ribotypes were collected (table 1).

**Table 1. Geographical distribution of species from dataset of this study.**

| Species | Geographic distribution |
|---|---|
| ***F.paulensis*** | (Brazil) Atlantic Ocean, Balearic Islands(Western Mediterranean Sea) |
| ***F.ruetzleri*** | Atlantic Ocean |
| ***F.yasumotoi*** | Australia, Japan (?) |
| ***G.australes*** | Pacific Islands, North Atlantic Sea (Canary Islands and Madeira Archipelago), Western Mediterranean Sea, China |
| ***G. balechii*** | Indonesian |
| ***G.belizeanus*** | Atlantic North(USA), Bahamas, Bermuda, Cancun Canary Islands, Eastern mediterranean Sea, Red Sea |
| ***G. caribaeus*** | Canary Islands |
| ***G.carolinianus*** | Canary Islands, Eastern Mediterranean Sea |
| ***G. cheloniae*** | Pacific Islands |
| ***G.excentricus*** | Canary Islands |
| ***G.honu*** | Pacific Islands |
| ***Gambierdiscus ribotype 2/ G.jejuensis*** | Japan |
| ***G. lapillus*** | Australia |
| ***G.polyniensis*** | Pacific Islands |
| ***G. pacificus*** | Pacific Islands |
| ***G. scabrosus*** | Japan |
| ***G. toxicus*** | Pacific Islands |
| ***Gambierdiscus type 4*** | *Pacific Islands* |
| ***Gambierdiscus type 6*** | Pacific Islands |

## 3.2 Genetic distances for *Gambierdiscus* and *Fukuyoa* genera

Genetic distances were obtained and saved in excel files for molecular markers LSU D8-D10 rDNA, although subsequent section will be explained main results only for *G.australes*. Genetic distances were interpreted in the trees.

## 3.3 Genetic distances for *G.australes.*

Genetic distances within *G. australes* species ranged between 0. and 0.021 for D8-D10 rDNA. Although, most of the strains genetic distances ranged between 0 and 0.002. These last distances were considered low and were not taken into account because they can be attributed to technical errors to get the sequences. However, KY448382 isolate VGO1258 from the Canary Islands has a higher genetic distance, which ranged between 0,0190 and 0.021. This isolate already was treated was treated as different ribotype of *G. australes*[24].

Results for D1-D3 region genetic distances were *G.australes* were similar, range was 0.002 and 0.03. Most of distances ranged between 0.0 and 0.001. However, are two strains one with code EF202970.1 (isolate RAV 92) is from Rairua, Raivavae Island, Australes Archipelago in the Pacific Ocean [25] , its genetic distance ranged between 0.003 and 0.006; and for the strain KY448417.1 (isolate VGO 1270)[26] from the Canary Islands) had the genetic distance between 0.003 and 0.004.

As a result of mantel test there was no correlation between genetic distance and the geographical distance p.value>0.05.

To sum up, distances until 0.02 were very small and could be an error in the obtention of the sequences, this could be for example an error of polymerase. Strains with hihg distance in other studies have been considered as *G.australes* but diffetent ribotypes. In addition, D8-D10 rDNA and D1-D3 rDNA were not considered a good marker to explain differences between geographical points, these results are in concordance with the phylogenetic trees that will be showed below.

### 3.4 Genetic diversity of *Gambierdiscus spp.*

For revealing haplotypes in *Gambierdiscus* and *Fukuyoa* genera, sequences of LSU D8-D10 were studied with DNAsp, results are summarized in (table 1).

Comparing haplotypes/phylotypes within species, in comparison of number of studied sequences, almost all species have high number of haplotypes (n°H). This phenomenom is showed with the haplotype gene diversity (H) as well.

However, *G. excentricus* has low quantity of haplotypes. Most of the sequences in the analyses are from the Canary Islands (North Atlantic Ocean) except two sequences from Brazil (South Atlantic)[27]. Low values of nucleotide diversity and low haplotype diversity could be a result of new colonization or a bottleneck. Level of diversification in the Canary Islands seems to be low, considering that analysis has showed only one haplotype for 15 sequences; this could be an indicator of recent introduction of microalgae in the Canary Islands. Nevertheless, more markers should be studied to understand if there could be a recent introduction.

For *G. australes* species are low as well, therefore could be a bias of the data because almost sequences are from the Canary and Balearic Islands. Moreover, the number of haplotypes of Balearic Islands are higher than the haplotypes of the Canary Islands. This could represent a different introductions of *G. australes* in the Balearic Islands or that have had more time to diversify being an old introduction with more time than the Canary Islands. *F. paulensis* from the Balearic Islands has two haplotypes, so it could be two different introductions or that taxa have already diverged.

Further analyses are needed in order to confirm these preliminary results. Results from different ribotypes and types has to analyse deeper. It is not clear in the articles with is the difference of type, ribotypes, etc.

| Populations | | N | S | S.pol | $\pi$ | nº H | H | Fu F's D statistic | Fu F's F statistic |
|---|---|---|---|---|---|---|---|---|---|
| **Pooled** | | 434 | 262 | 204 | 0,04831 | 113 | 0,914 | -10,59** | -7,086** |
| **G. scabrosus** | | 5 | 468 | 15 | 0,01709 | 5 | 1 | 0,812 | 0,8655 |
| **G. toxicus** | | 11 | 481 | 24 | 6,945 | 10 | 0,982 | -0,87 | 0,936 |
| **G. pacificus** | | 17 | 483 | 12 | 0,00335 | 6 | 0,588 | -2,419* | -2,661* |
| **G. lapillus** | | 4 | 482 | 0 | 0 | 1 | 0 | 0 | 0 |
| **Gambierdiscus type 6** | | 10 | 483 | 21 | 0,01256 | 6 | 0,844 | 0,42786 | -0,6542 |
| **G. balechii + Gambierdiscus type 6** | | 12 | 482 | 23 | 0,01229 | 7 | 0,00565 | -0,42292 | -0,6925 |
| **G.belizeanus** | | 12 | 390 | 46 | 0,02071 | 9 | 0,955 | -2,59** | 2,822** |
| **Gambierdiscus ribotype 2/ G.jejuensis** | | 11 | 483 | 12 | 0,00595 | 8 | 0,891 | -1,144 | -1,339 |
| **G. caribaeus** | | 44 | 483 | 20 | 0,00223 | 17 | 0,679 | -4,728* | -4,699* |
| **G.carolinianus** | | 16 | 466 | 13 | 0,00794 | 13 | 0,95 | -2,70353* | -2,85702* |
| **G.polyniensis** | | 6 | 555 | 17 | 0,01037 | 6 | 0,909 | 1,536** | 1,2486 |
| **G.australes** | | 111 | 401 | 23 | 0,00121 | 12 | 0,189 | -6,587** | -6,052** |
| **G.australes** (Balearic Islands) | | 29 | 435 | 15 | 0,00253 | 5 | 0,261 | -4,356** | -4,4262** |
| **G.australes** (Canary Islands) | | 57 | 426 | 10 | 0,00082 | 6 | 0,0045 | -4,924** | -4,797** |
| **G.excentricus** | | 43 | 405 | 4 | 0,00092 | 4 | 0,136 | -4,218** | -4,216** |
| **G.excentricus** (Canary Islands ) | | 15 | 405 | 0 | 0 | 1 | 0 | 0 | 0 |
| **Fukuyoa spp.** | | 24 | 424 | 75 | 0,0303 | 10 | 0,75 | 1,148 | 0,326 |
| **F.ruetzleri** | | 5 | 480 | 4 | 0,00333 | 4 | 0,9 | -1,093 | -1,113 |
| **F.paulensis** | | 9 | 449 | 8 | 0,00396 | 2 | 0,222 | -2,029* | -2,202* |

**Table 2. Genetic diversity of each specie and genus. N, number of strains; S, length of sequences; S.pol, number of segregating sites; nº H, number of haplotypes; π, nucleotide diversity ; H, haplotype diversity; *, p-valor < 0.05; **, p-valor < 0.02.**

## 3.5 Phylogenetic trees

### 3.5.1 Phylogenetic trees based on genetic distances

Phylogenetic trees with all Gambierdiscus and Fukuyoa species are huge, therefore, only phylogenetic trees based on *G.australes* sequences have been presented as a result of distance methods. Phylogenetic result with all species (*Gambierdiscus* and *Fukuyoa*) is pesented only by maximum likelihood method.

G. australes species is common in the Canary Islands, Madeira and the Balearic Islands. In order to see any relationship between populations from the Atlantic Ocean and the Mediterranean Sea, specific phylogenetic analyses have been performed. Analyses were based in two molecular markers D8-D10 rDNA and D1-D3 rDNA. Firstly, phylogenetic trees based on distances are showed. Previously with genetic distances matrix was possible to observe a little change in G. *australes* that most of them are not considered informative. Although, 3 strains with codes VGO 1248 (from Canary Islands), 17- 256 and 17- 216 (from the Balearic Islands (date in brown) are different of rest of G. australes species (Fig. 3). In the future 17-256 and 216 could be considered as new ribotypes**.**

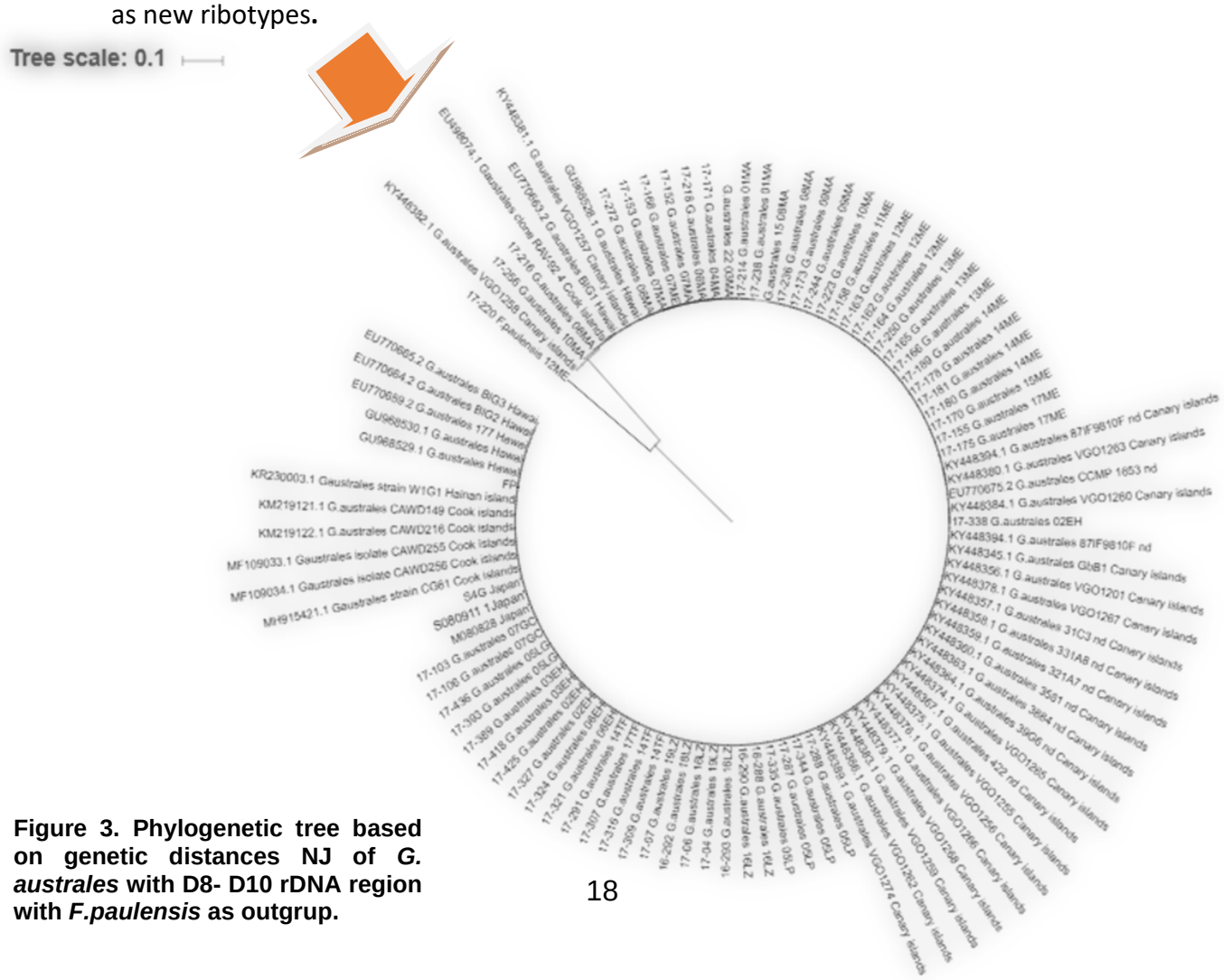



**Figure 3. Phylogenetic tree based on genetic distances NJ of *G. australes* with D8- D10 rDNA region with *F.paulensis* as outgrup.**

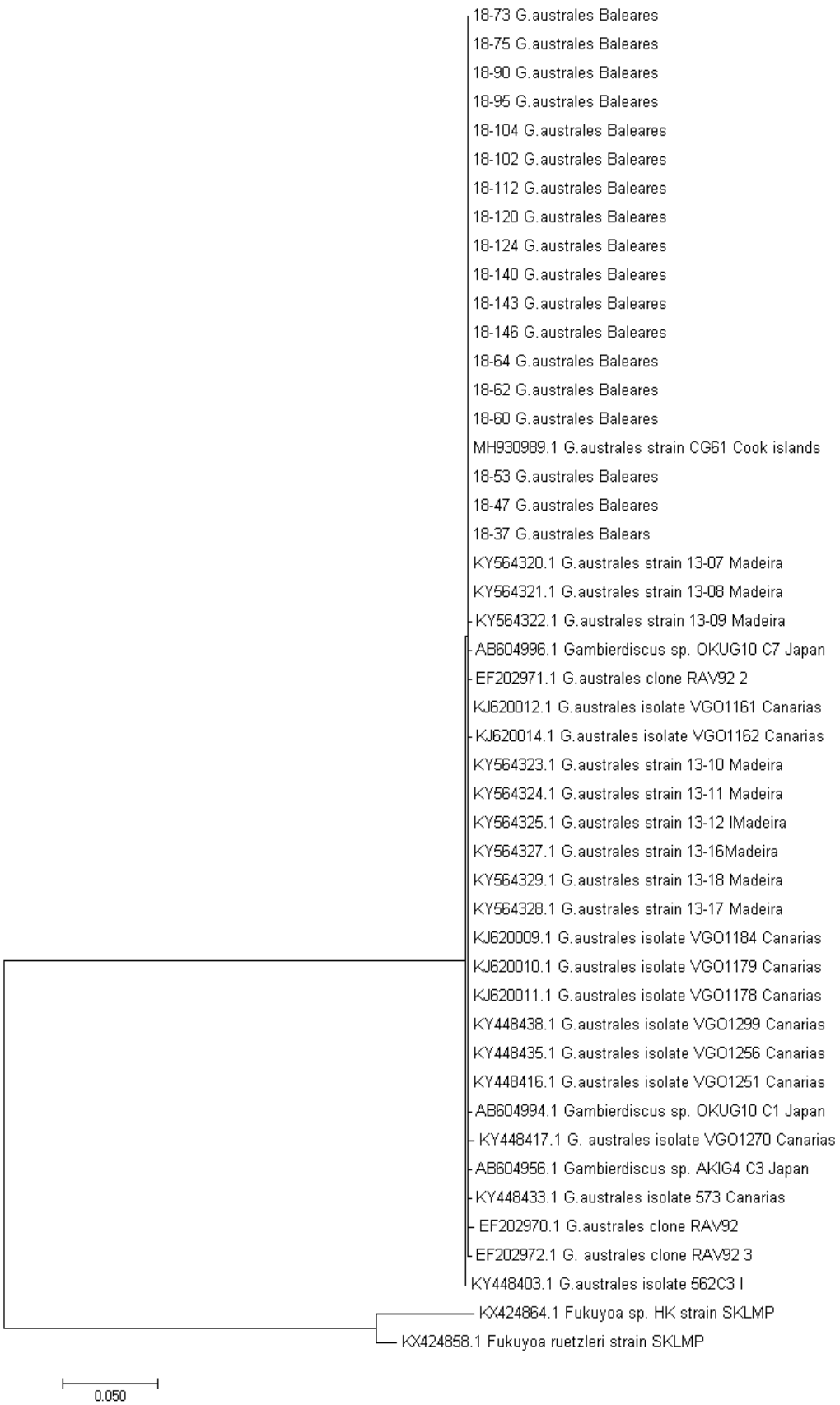



**Figure 4. Phylogenetic tree based on genetic distances NJ of G. australes with D1- D3 rDNA region with *F.ruetzleri* as outgrup.**

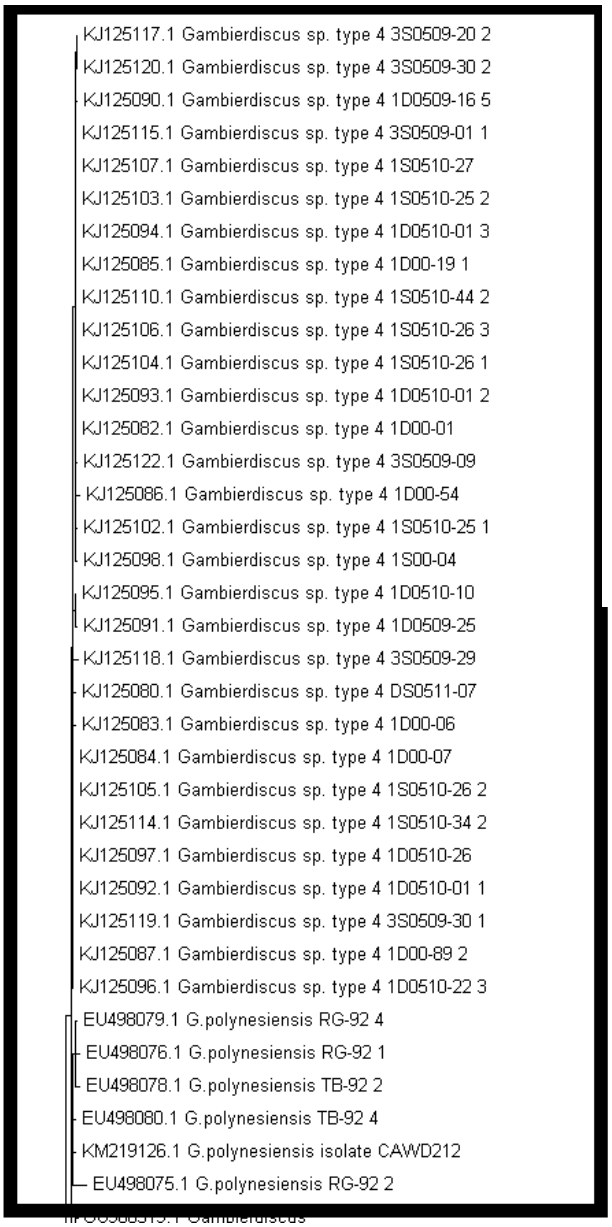
For D1-D3 marker in distance matrix was possible to observe some differences but, in the tree, based on distance, the differentiation is not possible to appreciate.

### 3.3.2 Phylogenetic trees based on Maximum likelihood for *Gambierdiscus* and *Fukuyoa* genera.

The evolutionary history was inferred with molecular marker LSU D8-D10 rDNA by using the Maximum Likelihood method based on the General Time Reversible model[28]. The tree with the highest log likelihood (-4379.53) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining (NJ) algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. A cause to obtain good align, some sequences were dropped and final align involved 396 nucleotide sequences. All positions containing gaps and missing data were eliminated. There was a total of 449 positions in the final dataset.

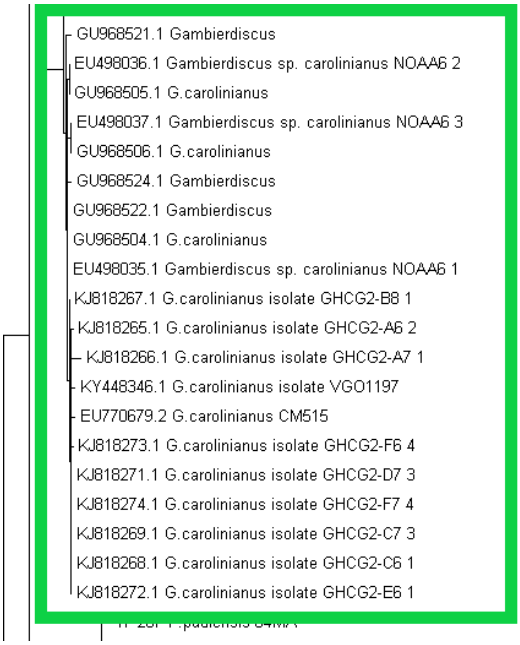
In general clades are well defined, although almost each clade has exceptions. *Fukuyoa* clade are divided in two: one part is next to *G. polynesiensis*, *G.silvae*, *Gambierdiscus ribotype 3* and *G. carolinianus* and the other with 2 strains labelled as *Fukuyoa yasumotoi* are close to *G.scabrosus* (Fig. 5). Differences between geographical points are not really present within species, and species from geographical points are placed together in some clades.

**5. Clades from D8-D10 of all data the maximum likelihood methods**

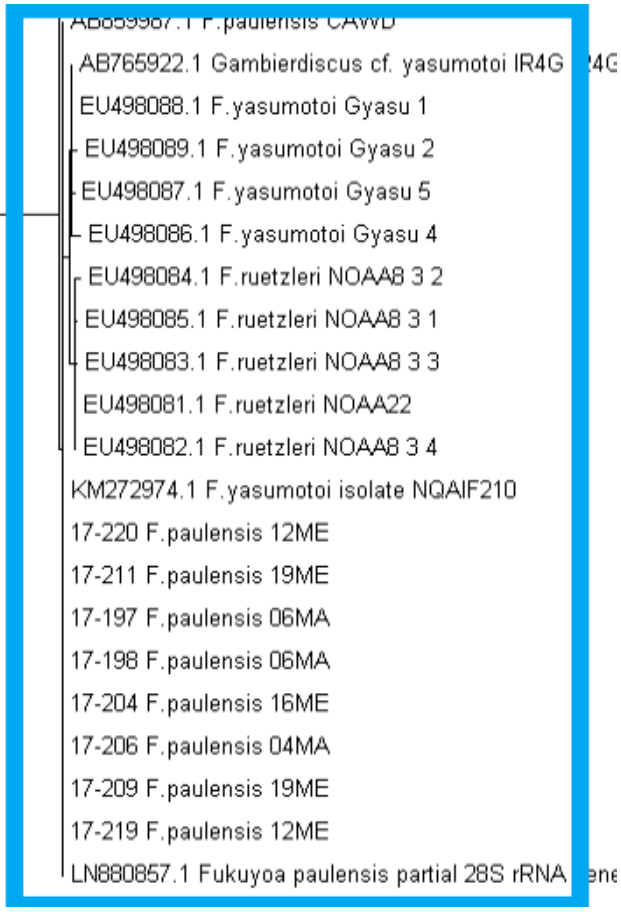



Clade *G. polynesiensis* and Gambierdiscus ribotype 4.

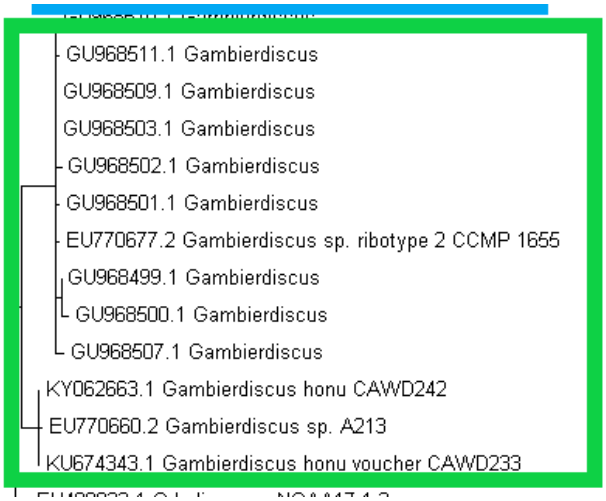

Clade *G. silvae* and Gambierdiscus ribotype 3.




Clade *G. carolinianus*

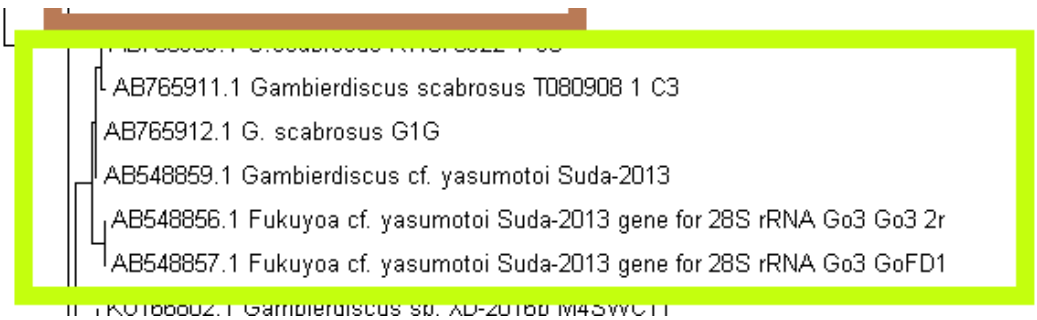



Clade *Fukuyoa* spp.

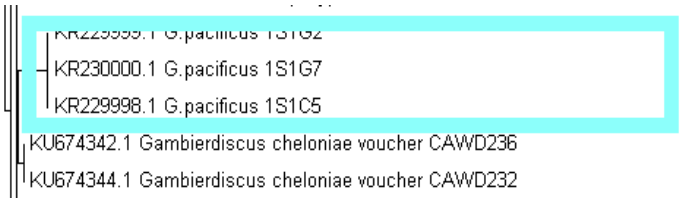



Clade *Gambierdiscus* ribotype 2 and *Gambierdiscus honu*

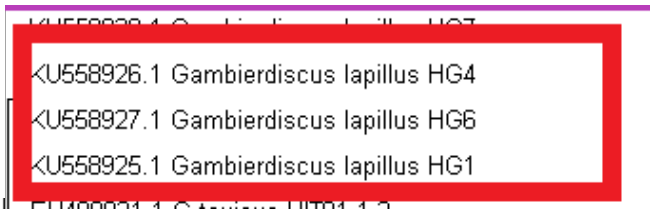

Clade *G.belizeanus.*




Clade *G.scabrosus* and *F. yasumotoi.*

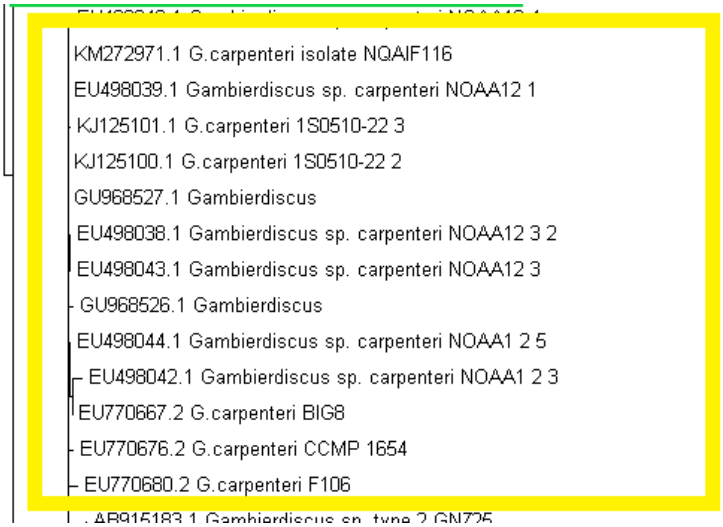



Clade *G. pacificus* and *G. cheloniae*

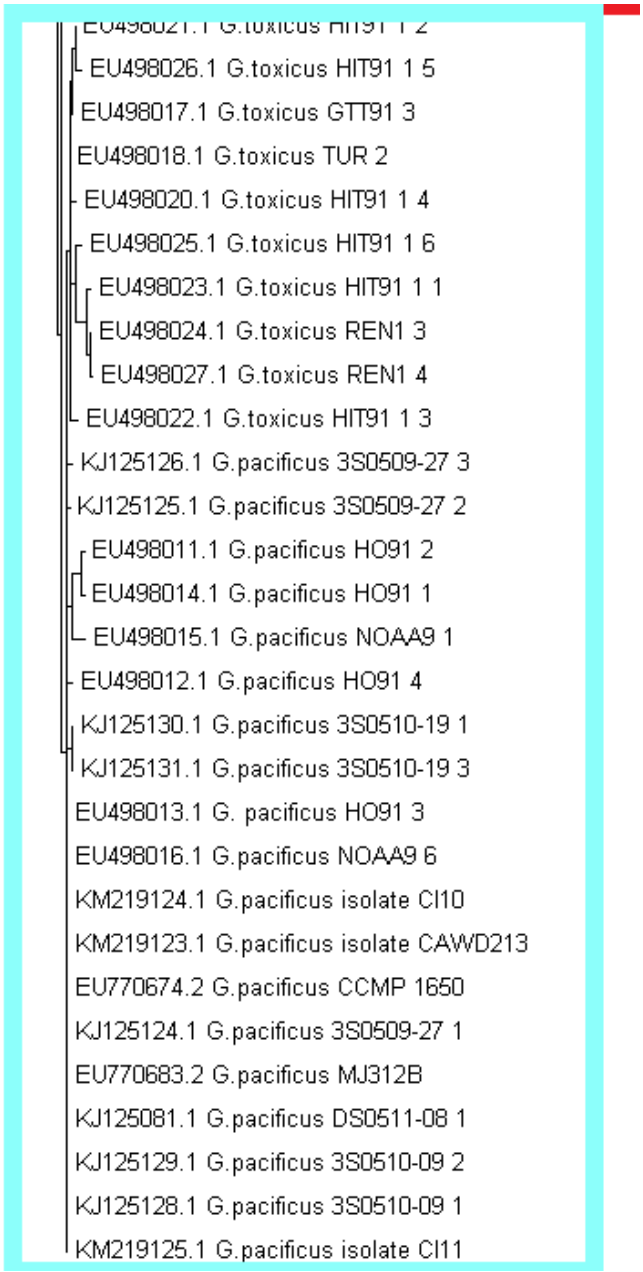



Clade *G. lapillus*

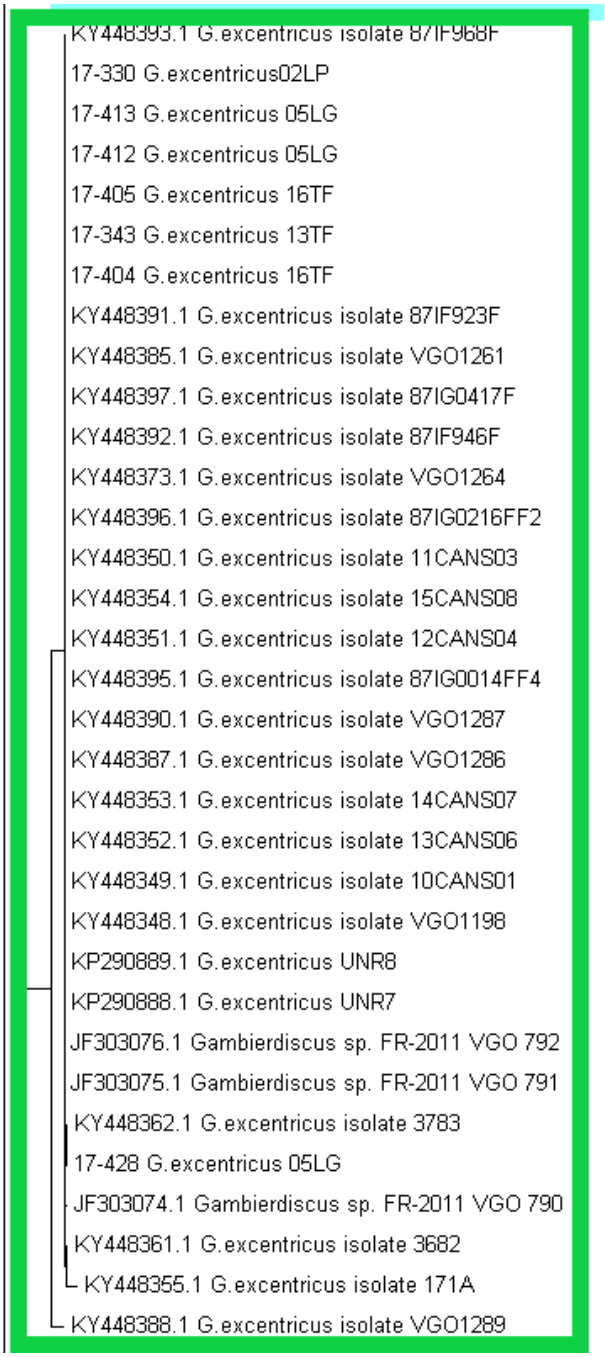

Clade *G. balechii* and *Gambierdiscus* ribotype 6



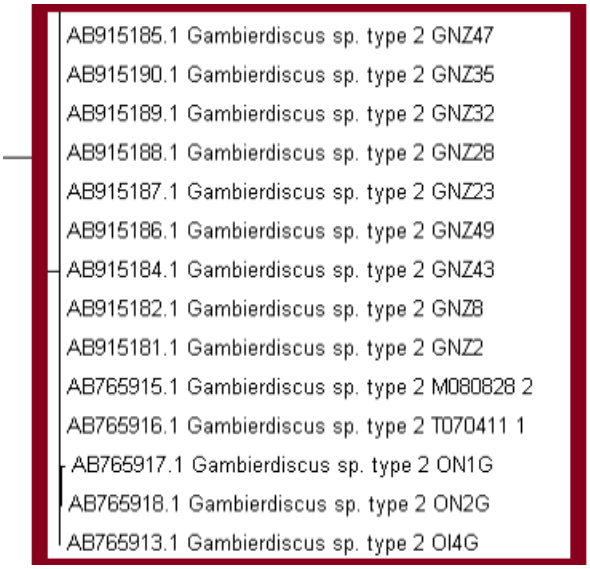

Clade *G. carpenteri*



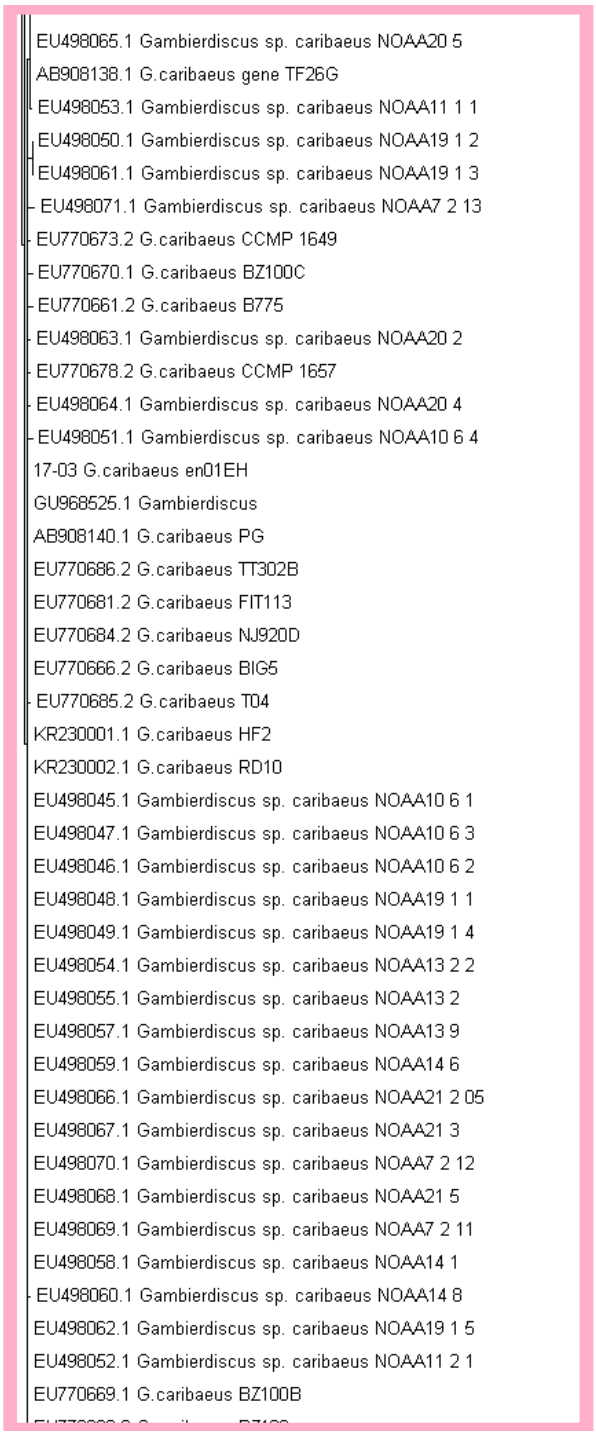

Clade *G. toxicus* and *G. pacificus*



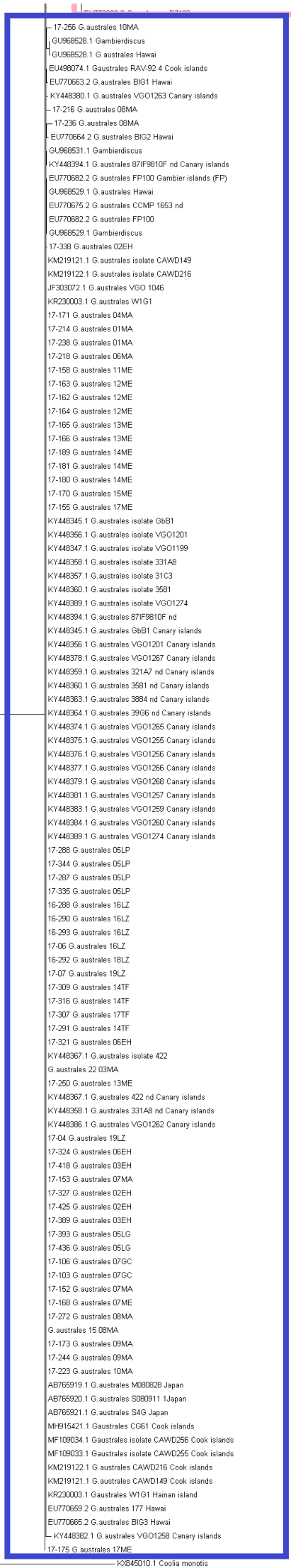

Clade *Gambierdiscus excentricus*




Clade *Gambierdiscus* type 2
or *G. jejuensis*




Clade *G. caribaeus*




Clade *G. australes.* See in next trees phylogenetics of *G. australes.*

# 4. Discussion

In the present study phylogeographical approach of all strains from Genbank was done with genetic markers that historically have been used to identify species[29–32]. These markers are not ideal to explain process of expansion range, but still some indications of processes could be present.

Large dataset with D8-D10 marker (n= 434, 592 pb) was created, final analyses contains at least 15 species and different ribotypes for *Gambierdiscu*s and 3 species of *Fukuyoa* genus.

As a result of genetic diversity analysis, a low presence of haplotypes can be observed for *G.australes* and *G. excentricus.* Many *G. australes* sequences are from Balearic and Canary Islands; most of genetic distances between strains are very small (0.002), so could be a recent introduction or could be a bias for molecular marker that is very conservative within species. In phylogenetic trees there are also not differences between geographical regions. Mantel test shows for G.*australes* that there is not genetic divergence between all strains. For each specie further studies have to be done in order to check the possible differences.

Tropical Pacific regions has the typical cases Ciguatera, and they present more *Gambierdiscus* species and high level of endemism, as well. If we check the origin of the species in the database we can see some species are cosmopolitan such as: *G.australes*, *G.belizeanus* and *F.paulensis* (table 1). These three species are also reported in the Mediterranean Sea, which is a warm-temperate area, far away tropical areas and where any feasible case of Ciguatera has not reported. Populations of Mediterranean Sea has to identify and more studies about population expansion are required to evaluate the risk of Ciguatera.

In reference to available information from the databases, for some strains that in GenBank are labelled as one species, in our phylogenetic tree, these strains are placed in different clades, for instance, *F. yasumotoi* and *Gambierdiscus ribotype 2*. Further revision in the literature is necessary to do it to update the databases, part of the job

of this work was a revision of taxonomy, but further revisions in each strains has to be analyse.

In this study the separation of the clades *Gambierdiscus* and *Fukuyoa* is not totally observed, but with D1-D3 rDNA and SSU markers *Gambierdiscus* and *Fukuyoa* from others articles, genera are located in different clades (fig. 6).

Sequences of ITS marker are used only to separate species which by morphology are very similar, and with classical markers are placed together[33]. Maybe will be a good marker to study geographical differences between species and to explain if there are processes of expansions.

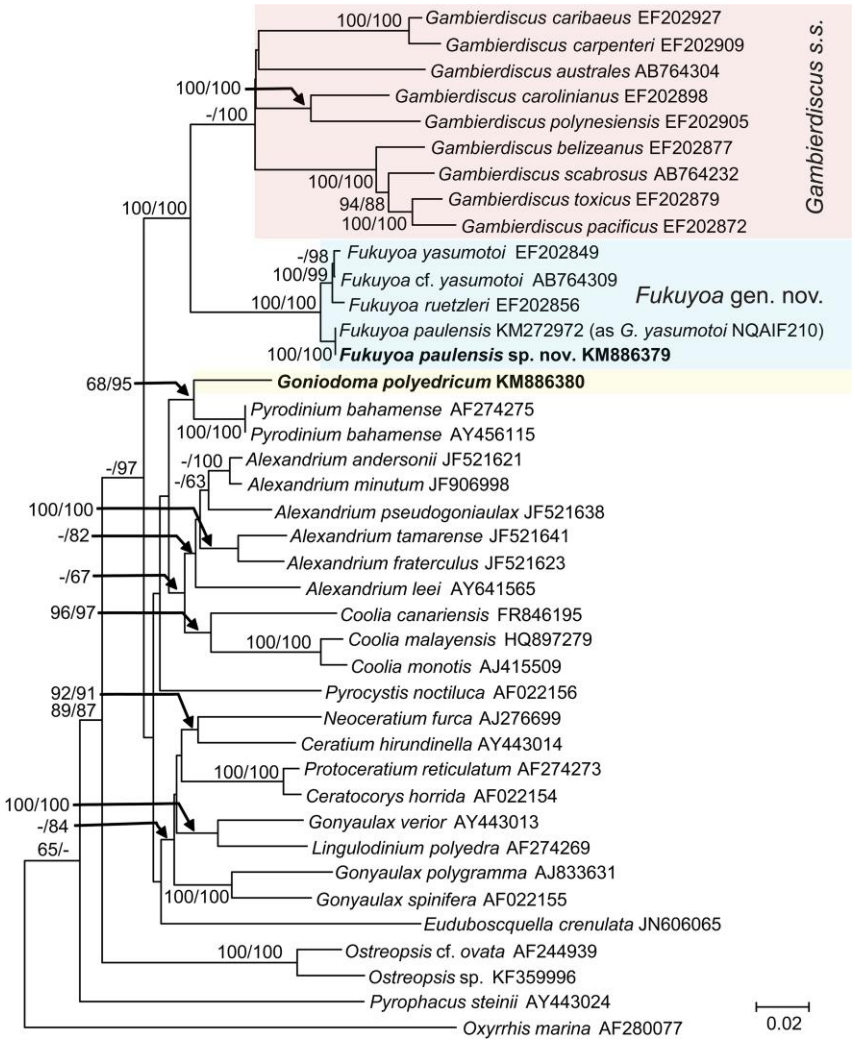



**Figure 6. Results of Gómez et al. 2015. SSU rDNA-based phylogeny of *Fukuyoa paulensis* gen. et sp. nov. and *Goniodoma polyedricum* with some gonyaulacoid dinoflagellates from Gómez et al. 2015. Sequences obtained in this study are bold-typed. Support of nodes is based on bootstrap values of ML/NJ with 1000 and 500 resamplings, respectively. Only values greater than 60 are shown. *Oxyrrhis marina* was used as outgroup.**

# Chapter 2

## Modelling absence or presence of *Gambierdiscus* and *Fukuyoa* genera

In the second part of this study, a model logistic based on the presence and absence of *Gambierdiscus* and *Fukuyoa* species was performed by R software. A revision of the literature has been done to compile all locations, where CTX-producers were reported. Locations were codified by latitude and longitude; subsequently environmental data of these locations were compiled by ArcGIS (ESRI 2011. ArcGIS). The model was based on presence and pseudoabsence. Pseudoabsence is an artificial data, that represents species which were not present in the sampling point[13]. To confirm absences in marine species is very difficult, particularly depth of the samplings could represent a bias. Normally sampling of *Gambierdiscus* and *Fukuyoa* is proceeded by apnoea a few meters of depth. In this study, has been used a multiple logistic regression, involving a logit link and binomial error distribution[16,17]. Logistic regression is one generalized linear model that is allow linear modelling when the response follow a non-normal distribution, besides is possible to work with binary variables (presence and absence)[18,34].

# 5.Methods

## 5.1 Extraction of environmental data (Annex 5)

In each sampling point where CTX-producers were reported and it was possible to find the coordinates, environmental data was downloaded by ArcGIS (ESRI 2011, CA. Environmental Systems Research Institute), from the database Bio-ORACLE v2.0 (http://www.bio-oracle.org/). Layers downloaded were: Surface, Benthic - Benthic - Minimum depth, Maximum depth Benthic - Average depth, Coral Reefs 2010 and Bathymetry. As a result, 315 rasters of environmental data were obtained.

Environmental data from Bio-Oracle contains information about (Annex 4):

- Currents velocity ($m^{-1}$)
- Ice thickness (m)

- Sea ice concentration (Fraction)
- Nitrate (mol.m$^{-3}$)
- Phosphate (mol.m$^{3}$)
- Silicate (mol.m$^{-3}$)
- Dissolved molecular oxygen (mol.m$^{-3}$)
- Iron (µmol.m$^{-3}$)
- Chlorophyll (mg.m$^{-3}$)
- Phytoplankton (µmol.m$^{-3}$)
- Primary productivity (g.m$^{-3}$.day$^{-1}$)
- Calcite (mol.m$^{-3}$)
- pH
- Photosynt. Avail. Radiation (E.m$^{2}$. day$^{1}$)
- Diffuse attenuation (m$^{-1}$)
- Cloud cover (%)
- Salinity

All *rasters* were save in a excel file with R software following instructions that are explained in Annex 5.

## 5.2 Selection of variables

### 5.2.1 Analysis of correlation (Annex 5)

A file with all downloaded environmental data from ArcGIS and coordinates was created as " POINTS.csv ". First observation of the data were performed and a general description of the data were obtained by str() and summary() (Annex 4). In addition, variables related to ice information were removed, and some binary variables were recodified by GIS. The recodification of binary variables was done in order to include the information contained in binary variables in the principal components analysis (PCA); for example, variable presence of corall, were recodified as the distance from corall. In all steps, biological interpretation was considered.

To reduce the dimensions of the dataset, a correlation matrix by Pearson method were estimated of environmental data. Matrix was calculated by function cor(), and excel file with correlation coefficients was saved.

## 5.3 Principal Component Analysis (Annex 6)

Before analysing interrelationships among the variables by Principal Component Analysis (PCA), binary variables were removed from the dataset. PCA was performed with selected variables from the previous matrix correlation. PCA were performed with and without standardization of data. Numerical range of the variables and units are different; therefore, the standardization of data was necessary (Annex 4).

PCA were calculated with prcomp() following these steps:

- PCA equations estimations
- Check the variability of components and the importance of each variable to the components.

## 5.3 Modelling (Annex 7)

In this study, classical logistic models based on generalized linear models of presence and pseudo-absence probability[17,29] were performed for all species of *Gambierdiscus* and *Fukuyoa* with "Biomod2" package of R software. Logistic models were performed for each species, and for each genus (*Gambierdiscus* or *Fukuyoa*) separately. After modelling, models were evaluated following the guideline of the study of Manel et al. 2001[18].

Original data set was reduced to 28 variables. Binary data of presence or absence of the taxon in each sampling point was created with Excel (v.1808) (fig.9). Presence was scored like 1, absence like 0 (Annex 7).

After creation the binary variables, models were performed following the next steps:

Step1:
-Create a matrix with all data, there is indications of each object involved in our model.

Step. 2:
-Proceeding of modelling our data

For modelling, generalized linear models "GLM" option was chosen data and was split in two subdata (testing data) and 3 runs of variables have been developed and evaluated. This option is specific for logistic regression with binary variable response

binomial. Model will be an equation to predict if in one point species or genus will be present or not.

As Alliouche 2006 well explain models generating are usually evaluated by comparing the predictions with a set of validation sites and constructing a confusion matrix that records the number of true positive (a), false positive (b), false negative (c) and true negative (d) cases.

The evaluation of the model was done with the same package (*Biomodels2*) and the function get_evaluations().

Measures of evaluation are True Skill Statistic (TSS) or Hanssen- kuipers discriminant that measure accuracy, Receiver operating characteristic (ROC) (Fielding and Bell 1997), sensitivity and specificity[37]. Sensitivity is the proportion of correctly predicted presences and specificity (the proportion of correctly predicted absence). The best model was chosen as model with the highest score of TSS, following indications of Allouche 2006[37].

| | | Validation data set | |
|---|---|---|---|
| | | Presence | Absence |
| Model | Presence | $a$ | $b$ |
| | Absence | $c$ | $d$ |

| Measure | Formula |
|---|---|
| Overall accuracy | $\frac{a+d}{n}$ |
| Sensitivity | $\frac{a}{a+c}$ |
| Specificity | $\frac{d}{b+d}$ |
| TSS | sensitivity + specificity – 1 |

**Figure 7. Equations of parameters to evaluate logistic models from Allouche 2006**.

# 6. Results

## 6.1 Extraction of environmental data

Finally, our set contained 311 marine environmental variables of 264 points where *Gambierdiscus* or *Fukuyoa species* had been found and it was possible to find the coordinates of the sampling points.

## 6.2 Analysis of correlation (Annex. 4)

Results of correlation were saved in excel files. As a result of the analysis, it was observed big correlation (scores range 0.90 to 1) between "lt.max, lt.min, max, min and average" of the types of environmental variables. For instance, benthic temperature can be characterized by maximum benthic temperature, minimum benthic temperature and average benthic temperature. and all these variables are high correlated. Therefore, only one variable of average of benthic temperature was left in the analysis. After correlation the data set was reduced to 33 environmental variables, for example Benthic Mean Depth (BMD) Nitrate and BMD Phosphate are high correlated.

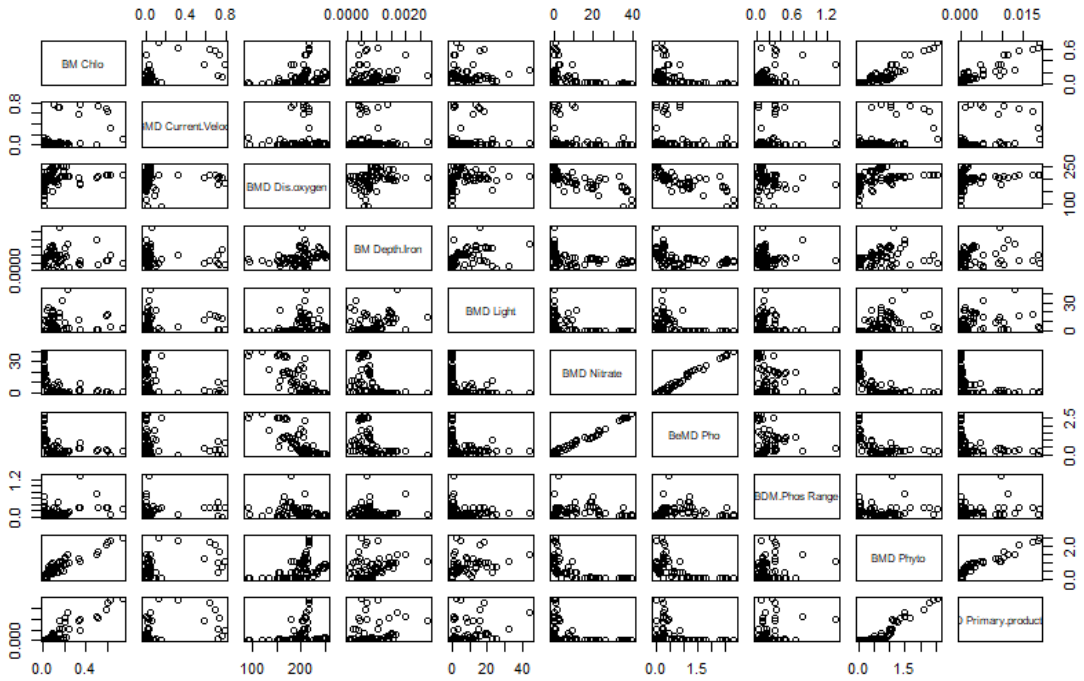



**Figure 8. Descriptive plot of relations of environmental variables.**

## 6.3 Principal Component Analysis

In this manuscript is only showed the PCA with standardized data, process are explained in annex 6.

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     3.3378 2.8303 1.84345 1.5482 1.50756 1.13869 1.00756 0.94033 0.86846 0.71330
Proportion of Variance 0.3277 0.2356 0.09995 0.0705 0.06685 0.03814 0.02986 0.02601 0.02218 0.01496
Cumulative Proportion  0.3277 0.5633 0.66323 0.7337 0.80058 0.83872 0.86857 0.89458 0.91676 0.93173
                         PC15    PC16    PC17    PC18    PC19    PC20    PC21    PC22    PC23     PC2
Standard deviation     0.4285 0.39577 0.35610 0.34096 0.29336 0.28344 0.23876 0.20862 0.16084 0.1414
Proportion of Variance 0.0054 0.00461 0.00373 0.00342 0.00253 0.00236 0.00168 0.00128 0.00076 0.0005
Cumulative Proportion  0.9775 0.98206 0.98579 0.98921 0.99174 0.99410 0.99578 0.99706 0.99782 0.9984
```

**Figure 9. Results of R from PCA with standardization**

Without standardization, the first principal component with the first component was possible to have 99% of variability, but if we observed the coefficients of each variable, for all coefficients of variables are 0.00 except the bathymetry that is 1 (fig. 9, 10).
For the second principal component, equation was based on dissolved oxygen (+) and silicates (-) and distance of coral presence (+), and not is based on bathymetry. The variance was plotted of the analysis, almost all variance is from the first component and thus was due to bathymetry (fig. 10)

Bathymetry was having all weight in the first principal component In contrast to without standarization, with standarization the proportion of variability explained was spread with almost all of variables and coefficient of bathymetry is 0.11. From the coefficients are showed previously, important variables are surface dissolved oxygen (+) as a positive variable, current velocity range (-) as a negative variable, nitrats (-) as negative variable, surface temperature (+) as a positive variable. By biplot, the presence of CTX-producers seems that are linked to places with high dissolved oxygen, high surface temperatures, oligotrophic and with low currents.

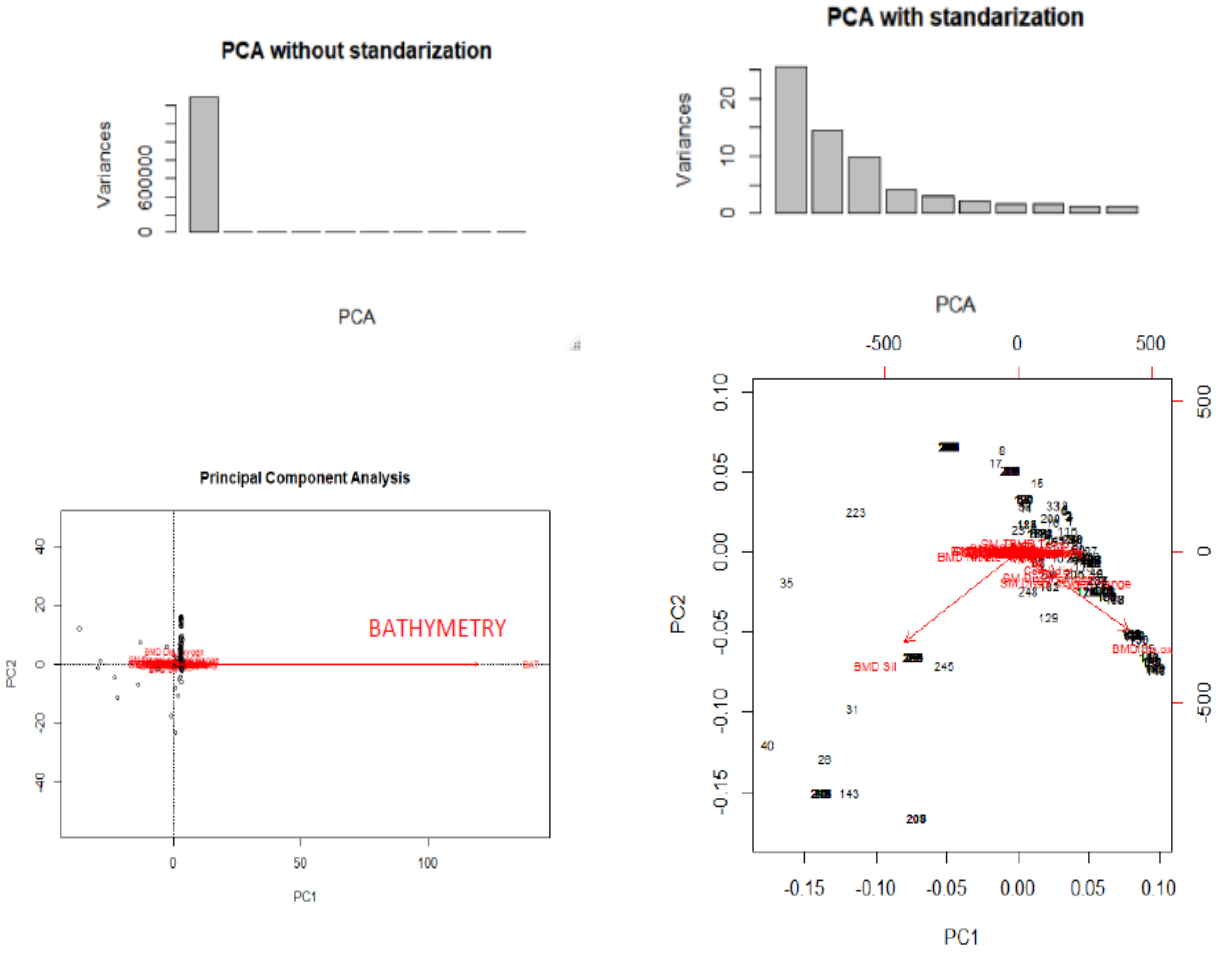



**Figure 10. Variance of PCA (both sides-up), Biplot PCA of two first principal components (down): not standarization (left), standarizated (right).**

## 6.4 Modelling distribution for each species

In this study, classical logistic models based on generalized linear models, as response variables was binary variable presence or pseudo-absence. These models were performed for all species of *Gambierdiscu*s and *Fukuyoa* with "Biomod"2 package of R software. Logistic models were performed for each specie, and for each genus (*Gambierdiscus* or *Fukuyoa*) separately.

## 6.5 Evaluation of Models

The best model was chosen as model with the highest score of True Skill Statistic (TSS) or Hanssen- kuipers discriminant, following indications of Allouche 2006[37]. TSS values are comprised between -1 to 1; when values are closed to 1 the better model is. For G*. cheloniae, G.toxicus, G.balechii*, *G. polynesiensis*, *G. silvae* we could not find the coordinates or species was present only in one sampling point, therefore was not possible to estimate the model. Moreover, for *G. carolinianus, G.* cf. *yasumotoi (F*. cf. *yasumotoi)* and *G. scabrosus* was not possible to find a model with parameters, run not converged.

Results of models are compiled in annex 8, in results were showed the formulas of the best model, the evaluation of models and the importance of the variables within models. Importance of variables are showed as well in table 3. If we see the results of TSS of the models that have been obtained, is easy to see that there are high values of TSS. It suggests that the achieved models are good models to predict the presence or absence of *Gambierdiscus* and *Fukuyoa* genera.

If we see in the table 3, the value to assess the importance of variables range 0 to 1 and is the relative number of times that variable has importance to the model. Values higher than 0.5 are coloured in red, which means that these variables have been appeared in 50% of the generated models. There are 3 variables that have importance values>0.5 and are in common in some variables:

- Variable 14: surface. PAR. mean (*G. belizeanus*, *Gambierdiscus* ribotype 4, *G. carpenteri* and Gambierdiscus ribotype 5).
- Variable 21: Surface silicates Mean (*G. excentricus*, *Gambierdiscus* ribotype 1, *Gambierdiscus* ribotype 2).
- Variable 22: Surface temperature Mean (*G. australes*, *G. excentricus*, *F. paulensis*)

It is strange that for all sequences together of *Gambierdiscus*, the surface dissolved oxygen is important and has presence in all models for *Gambierdiscus* analysis. But, in the analysis when species are modelling separately this variable has low importance.
For *Fukuyoa* genus in global seem to be also important the variable surface dissolved oxygen, although as *Gambierdiscus* genus, when is evaluate separately this not seems to be important for each *Fukuyoa* species.

These results are in concordance to the PCA, the two principal variables contained dissolved oxygen (+), nitrats, silicates (-) and surface velocity (-).

In general, we can conclude that *Gambierdiscus* and *Fukuyoa* species are reported in oligotrophic environments and with slow currents but with high dissolved oxygen.

| Species | Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G.belizeanus** | yes | | | | | | | | | 0.115 | | | | | 0.944 | | | | | | | | | 0.502 |
| **G.caribaeus** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **G.pacificus** | yes | 0.435 | | | | | | | 0.413 | | 0.197 | 0.196 | | | | | | 0.562 | 0.929 | | | | | |
| **G.excentricus** | yes | | | | | | | 0.095 | 0.313 | | | | | | | | | | | | 0.261 | 0.634 | 0.226 | |
| **Gambierdiscus_cf._yasumotoi** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **Gambierdiscus ribotype.1** | yes | 0.280 | 0.008 | | | | | | | | | | | | | | | | | | | 0.954 | | |
| **G.carpenteri** | yes | | | | 0.857 | 0.640 | | | | | | | | | 0.272 | | | | | | | | | |
| **Gambierdiscus.type.4** | yes | | | | | | | | | | | | 0.198 | | 1.000 | | | | | | | | | 0.328 |
| **G.australes** | yes | | | | 0.087 | 0.287 | | 0.120 | 0.050 | 0,205 | 0.162 | | | | | 0.380 | | 0.270 | | | | | 0.983 | 0.490 |
| **G.scabrosus** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **Gambierdiscus.ribotype.2** | yes | 0.280 | 0.008 | | | | | | | | | | | | | | | | | | | 0.954 | | |
| **G.polynesiensis** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **G.carolinianus** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **F.paulensis** | yes | | | | | | 0.868 | 0.301 | | | 0.319 | | | | | | | | | | | | 0.687 | 0.474 |
| **G. balechii** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **G.honu** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **G.silvae** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **G.lapillus** | yes | | | | | | | | | | | | | 0.772 | | | | 1.000 | | | | | | |
| **G.cheloniae** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **Gambierdiscus.type.6** | yes | | | 0.963 | | | 0.086 | | | | | | | | | | | | | | | | | |
| **G.toxicus** | no | | | | | | | | | | | | | | | | | | | | | | | |
| **Gambierdiscus.type.5** | yes | | | | | | | | | | | | | | 0.602 | | | | | | | | | |
| **Gambierdiscus** | yes | | | | | 0.066 | | 0.102 | 0.034 | | 0.029 | | 1.000 | | | 0.202 | 0.225 | | | | | | 0.082 | |
| **Fukuyoa** | yes | | | | | | | | 0.251 | | 0.452 | | 0.976 | 0.449 | 0.654 | 0.684 | 0.644 | | | 0.583 | | | | |

**Table 3. Importance of of variables of logistic models for each species and genus. (in red high values >0.5**

1 BMD Iron.Mean

2 BMD .Phosphate.Range

3 BMD .Phosphate. Mean

4 BMD.Salinity.Mean

5 BMD.Silicate.Range

6 BMD.Silicate.Mean

7 BMD.Prim.prod.Mean

8 BMD.Light.bottom.Mean

9 BMD Chloll.Mean

10 Surface_Calcite.Mean

11 Surface_Chlol.Min

12 Surface_dissolved_oxygen_range

13 Surface diffuse att. Range

14 Surface_Par.Mean

15 Surface_Pho_Mean

16 Surface_Current veloc_Mean

17 Surface_pH

18 Coral_Distance

19 Surface_Phyto.Mean

20 Surface_Prim.productivity.Mean

21 Surface_Silicate.Mean

22 Surface_Temperature.Mean

23 GEBCO_Mean.Bathymetry

# 7. Conclusions

General conclusions from the data:

There are not many microalgal studies about populations and expansion distribution. This study presents a preliminary approach to analyse expansion of CTX-producers, although classical markers are conservative and the study could not arrive at this point, with our results is not possible to distinguish strains from different geographical points. Some differences could appreciate but more analyses have to be done. Specific goal to understand the relationships between *G.australes* from the North Atlantic Sea and the Mediterranean Sea have not been achieved, because markers from databases are not adequate. Some differences could be appreciated but not important to check the differences between populations from the Atlantic and Mediterranean Sea. In the literature, is not possible to find many works of populations adequate markers to work with populations could be microsatellites, but in algae microsatellites have not been developed largely.

The PCA shows a tendency of environmental conditions for the presence of all CTX-producers, but important variables each models of each species are different.

General conclusions from the project:

In the first chapter, I learned how genetic analysis is performed. Basics to work with genetic data of populations. I learned more how to work and visualize large matrix of data. At first, it has been very difficult to manage large dataset, large alignments, and large matrices but after I have been more confident on it. Analysis in this work have been tried to do totally in R software and see how packages work in this type of data, but total analysis in R not always have been possible. For example, MEGA7 was used to check the alignments, for me MEGA7 has been more useful to visualize alignments. Therefore, classical genetic programs are necessary and sometimes. Another example

was the program DNAsp that I find easier to manage large datasets than “adegenet” R package.

In the second chapter, I learned again to work with large data. How to work in the binary response variables. To convert logistic variables to numerical variables (for example environmental variable: presence or absence of coral) to introduce this information to dataset to see correlations, this variable was codified as distance to coral skull.

I learned how to work the logistic model (new model for me) and I applied to ecological problems.

I think this work is a previous work for further geographical analysis, more complex but more informative such as maxent models. With Maxent models with shapes files from GIS is possible to have geographical maps based on probabilities of presence of species (Phillip et al. 2010), then could be more realistic than the logistic models.

To work in a continuous updating of the dataset have been not a good idea, because loads of analysis have been done. However, issue was very interesting, and I would manage IRTA data which compiles information of *Gambierdiscus* and *Fukuyoa* strains from Europe.

I think I have done big efforts to understand new concepts and new methods, that I have less time to discuss in depth the results. Anyway, results have to take as a previous work for future analysis.

For me, this type of final project was new experience and the methodology of specify objectives and tasks for each objective have contributed positively to organize and evaluate how project is going in all steps.

# 8. Glosary

BMD: Benthic Mean Depth
CFP: Ciguatera Food Poisoning
CTX: ciguatoxinas
GIS: geographical information system
IRTA: Institut de Recerca i tecnologia agroalimentàries
NJ: neighbour joining
PCA: Principal Component Analysis
TSS: True Skill Statistic
UPGMA: unweighted pair group method with arithmetic mean

# 9. References


1. Yasumoto, T. *et al.* Environmental Studies on a Toxic Dinoflagellate Responsible for Ciguatera. *Nippon SUISAN GAKKAISHI* (1980). doi:10.2331/suisan.46.1397
2. Lewis, R. J. & Holmes, M. J. Origin and transfer of toxins involved in ciguatera. *Comparative Biochemistry and Physiology. Part C: Comparative* (1993). doi:10.1016/0742-8413(93)90217-9
3. Friedman, M. A. *et al.* An updated review of ciguatera fish poisoning: Clinical, epidemiological, environmental, and public health management. *Marine Drugs* (2017). doi:10.3390/md15030072
4. Kumar-Roiné, S., Matsui, M., Pauillac, S. & Laurent, D. Ciguatera fish poisoning and other seafood intoxication syndromes: A revisit and a review of the existing treatments employed in ciguatera fish poisoning. *South Pacific J. Nat. Appl. Sci.* (2011). doi:10.1071/sp10001
5. Diogène, J. *et al.* Identification of ciguatoxins in a shark involved in a fatal food poisoning in the Indian Ocean. *Sci. Rep.* (2017). doi:10.1038/s41598-017-08682-8
6. Tester, P. A., Feldman, R. L., Nau, A. W., Kibler, S. R. & Wayne Litaker, R. Ciguatera fish poisoning and sea surface temperatures in the Caribbean Sea and the West Indies. *Toxicon* (2010). doi:10.1016/j.toxicon.2010.02.026
7. Gómez, F., Qiu, D., Lopes, R. M. & Lin, S. Fukuyoa paulensis gen. et sp. nov., a new genus for the globular species of the dinoflagellate Gambierdiscus (Dinophyceae). *PLoS One* (2015). doi:10.1371/journal.pone.0119676
8. Rhodes, L. L. *et al.* The epiphytic genus Gambierdiscus (Dinophyceae) in the Kermadec Islands and Zealandia regions of the southwestern Pacific and the associated risk of ciguatera fish poisoning. *Mar. Drugs* (2017). doi:10.3390/md15070219
9. Jang, S. H., Jeong, H. J. & Yoo, Y. Du. Gambierdiscus jejuensis sp. nov., an epiphytic dinoflagellate from the waters of Jeju Island, Korea, effect of temperature on the growth, and its global distribution. *Harmful Algae*

(2018). doi:10.1016/j.hal.2018.11.007

10. Reverté, L. *et al.* Assessment of cytotoxicity in ten strains of Gambierdiscus australes from Macaronesian Islands by neuro-2a cell-based assays. *J. Appl. Phycol.* (2018). doi:10.1007/s10811-018-1456-8
11. Pisapia, F. *et al.* Toxicity screening of 13 Gambierdiscus strains using neuro-2a and erythrocyte lysis bioassays. *Harmful Algae* (2017). doi:10.1016/j.hal.2017.02.005
12. PETERSON, A. T. & VIEGLAIS, D. A. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. *Bioscience* (2006). doi:10.1641/0006-3568(2001)051[0363:psiuen]2.0.co;2
13. Leathwick, J. R., Elith, J. & Hastie, T. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecol. Modell.* (2006). doi:10.1016/j.ecolmodel.2006.05.022
14. Phillips, S. “A Brief Tutorial on Maxent” in Species Distribution Modeling for Educators and Practitioners. *Lessons Conserv.* (2010).
15. Morales S, N. Modelos de distribución de especies: Software Maxent y sus aplicaciones en Conservación. *Rev. Conserv. Ambient.* (2012).
16. McCullagh & Nelder, P. *Generalized linear models*. (1989).
17. Jongman, R. H. G., Ter Braak, C. J. F. & Van Tongeren, O. F. R. *Data Analysis in Community and Landscape Ecology*. *Cambridge University Press* (1995).
18. Manel, S., Ceri Williams, H. & Ormerod, S. J. Evaluating presence-absence models in ecology: The need to account for prevalence. *J. Appl. Ecol.* (2001). doi:10.1046/j.1365-2664.2001.00647.x
19. R Core Team. R: A language and environment for statistical computing. http://www.R-project.org/. *R Foundation for Statistical Computing, Vienna, Austria* (2017).
20. Hall, T. A. BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/ NT. *Nucleic Acids Symp. Ser.* (1999).
21. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressiv... [Nucleic Acids Res. 1994] - PubMed result.

*Nucleic Acids Res.* (1994).

22. Librado, P. & Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp187
23. Fu, Y. X. & Li, W. H. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* (1997). doi:10.1093/oxfordjournals.molbev.a025753
24. Rodríguez, F. *et al.* "Canary Islands (NE Atlantic) as a biodiversity 'hotspot' of Gambierdiscus: Implications for future trends of ciguatera in the area". *Harmful Algae* (2017). doi:10.1016/j.hal.2017.06.009
25. Kibler, S. R., Tester, P. A., Kunkel, K. E., Moore, S. K. & Litaker, R. W. Effects of ocean warming on growth and distribution of dinoflagellates associated with ciguatera fish poisoning in the Caribbean. *Ecol. Modell.* (2015). doi:10.1016/j.ecolmodel.2015.08.020
26. Rodríguez, F. *et al.* "Canary Islands (NE Atlantic) as a biodiversity 'hotspot' of Gambierdiscus: Implications for future trends of ciguatera in the area". *Harmful Algae* **67**, 131–143 (2017).
27. Nascimento, S. M., Melo, G., Salgueiro, F., Diniz, B. dos S. & Fraga, S. Morphology of Gambierdiscus excentricus (Dinophyceae) with emphasis on sulcal plates . *Phycologia* (2015). doi:10.2216/15-61.1
28. Nei, M. & Saitou, N. The neighbor-joining method: a new method for reco... [Mol Biol Evol. 1987] - PubMed result. *Mol Biol Evol* (1987).
29. Litaker, R. W. *et al.* Global distribution of ciguatera causing dinoflagellates in the genus Gambierdiscus. *Toxicon* (2010). doi:10.1016/j.toxicon.2010.05.017
30. Xu, Y. *et al.* Distribution, abundance and diversity of Gambierdiscus spp. from a ciguatera-endemic area in Marakei, Republic of Kiribati. *Harmful Algae* (2014). doi:10.1016/j.hal.2014.02.007
31. Richlen, M. L., Morton, S. L., Barber, P. H. & Lobel, P. S. Phylogeography, morphological variation and taxonomy of the toxic dinoflagellate Gambierdiscus toxicus (Dinophyceae). *Harmful Algae* (2008). doi:10.1016/j.hal.2007.12.020
32. Chinain, M., Faust, M. A. & Pauillac, S. Morphology and molecular analyses of three toxic species of Gambierdiscus (Dinophyceae): G-

pacificus, sp nov., G-australes, sp nov., and G-polynesiensis, sp nov. *J. Phycol.* (1999). doi:10.1046/j.1529-8817.1999.3561282.x

33. Nishimura, T. *et al.* Genetic Diversity and Distribution of the Ciguatera-Causing Dinoflagellate Gambierdiscus spp. (Dinophyceae) in Coastal Areas of Japan. *PLoS One* (2013). doi:10.1371/journal.pone.0060882
34. Regression, L. Logit Models for Binary Data. *Bernoulli* (1978).
35. Warton, D. I. & Shepherd, L. C. Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. *Ann. Appl. Stat.* (2010). doi:10.1214/10-AOAS331
36. Gu, W. & Swihart, R. K. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biol. Conserv.* (2004). doi:10.1016/S0006-3207(03)00190-3
37. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* (2006). doi:10.1111/j.1365-2664.2006.01214.x

# 10. Annexes

## Annex1. Creation dataset

### 1.1 Obtaining sequences:

downloading Genbank sequences with codes from articles (when taxonomy in the literature has changed but, labels in the dataset are not updated. `Example Fukuyoa strains from Larsson et al.2019.`

```
Install.packages(seqir)
library(seqir)
fpaulensis<-c("KM272974", "MH312005", "LN880857", "AB859987", "EU498082",
"EU498081","EU498085", "EU498084", "EU498083")

seqfpaulensis<-read.GenBank(fpaulensis, seq.names = access.nb, species.names =
TRUE,gene.names = FALSE, as.character = TRUE)

seqfpaulensis<-as.matrix(seqfpaulensis)

#we save a fasta file with all sequences seqir.


write.fasta(seqfpaulensis,as.string=FALSE,names=fpaulensis,
file.out="fpau.fas")
```

# we combine all sequences from IRTA and from Ge file “A.fas”.

### 1.2 Obtaining aligments:

```
# open file(con el paquete “ape”)
Install.packages(ape)
library(ape)
A<-read.FASTA("C:/Users/Angi/Documents/Rmaster/A.fas", type="DNA")
summary(A)
class(A)

# align the sequences with Clustalw and “ape” package:
clustal(A, exec="clustalw2", pw.gapopen = 10, pw.gapext = 0.1)


#we open the file and check the alignment:

dnasetb<-read.FASTA("C:/Users/Angi/Documents/Rmaster/dnasetb2.fas",
type="DNA")
```

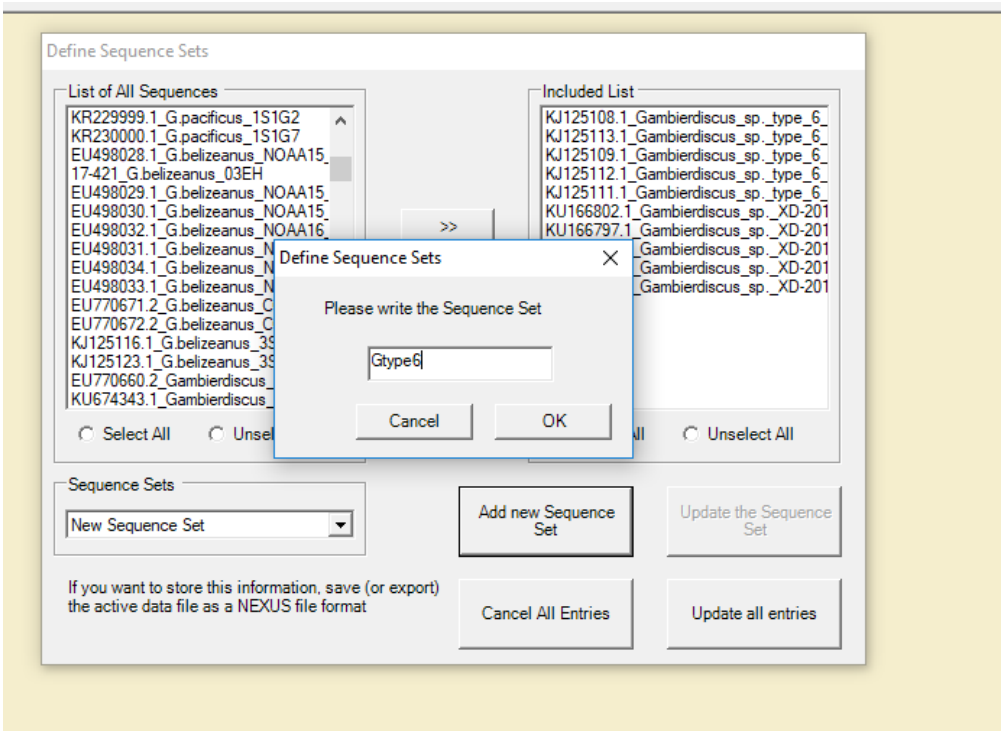
```
# as a result of aligmnet labels were changed by clustalW as “id”. Therefore to edit labels we extract dataset to matrix to change labels easily,after with MEGA7 aligment is checked manually and cut.

dnaset<-read.FASTA("C:/Users/Angi/Documents/Rmaster/align_tallat.fas", type="DNA") #se importan con las secuencias que previamente se han cortado con el MEGA7

#we extract the labels from sequences
names.txt <- read.delim("namesalign.txt", header = FALSE, sep = "\t")
head(names.txt)
align2<-as.matrix(dnaset)
rownames(align2)
names.txt<-as.matrix(names.txt) #we change the labels

#we create a new fasta file with all sequences and short sequences.
rownames(dnasetb)<-names.txt
```

## Annex 2. Analysis of genetic diversity

### 2.1 Instrucctions for DNAsp (analysis of genetic diversity)

Before the analysis, a data set was created, that contained all the possible sequences with the maximum possible length. For this reason, sequences were less than <565 bp were rejected. The analysis of genetic diversity was done through the DNAsp[22] program. To work with the DNAsp, you need the mega format file without gaps and without ambiguous positions, so the fasta alignment was converted to mega format with the MEGA7 converter, with the option of removing gaps and ambiguous positions.

After the conversion, the mega file was opened with the DNAsp with the option *File / Open unphase / genotip data file* and the subsets of populations were defined with the option: *Data / define sequence sets.* To establish the genetic subgrups, GenBank sequences which were labelled as *Gambierdiscus* sp were not taken.



## 2.2 Mantel test

### Step1: Estimate geographical distances

```
install.packages("geosphere")
library(geosphere)
data1<-read.csv("C:/Users/Angi/Desktop/TEMP/DATABASE/GIS    POINTS    presencia
absencia.csv", header=TRUE, sep=";", stringsAsFactors = FALSE)

attach(data1)


distGeo(p1     =     data.frame(data1$longitud,     data1$latitud),     p2     =
data.frame(data1$longitud, data1$latitud), a = 6378137, f = 1/298.257223563)


DISTANCIES <- NULL
for(P in 1:length(data1$latitud)){

           DIST     <-     distGeo(p1     =     data.frame(data1$longitud[P],
data1$latitud[P]),   p2   =   data.frame(data1$longitud,   data1$latitud),   a   =
6378137, f = 1/298.257223563)
           DISTANCIES <- cbind(DISTANCIES, DIST)

}

 colnames(DISTANCIES) <- c(1:length(data1$latitud))
 write.csv(x = DISTANCIES, file = "Distàncies.csv")
 DISTANCIES
```

### Step2: Genetic distances

```
install.packages("ape")
install.packages("phangorn")
library(ape)
library(phangorn)
align<-read.FASTA("C:/Users/atudo/Desktop/TEMP/DATABASE/PROVAGEN.fas",
type="DNA")

dnaphy<-as.phyDat(align) # change to format phydata
distprova<-dist.hamming(dnaphy)
head(distprova)
prova4<-as.matrix(distprova) # create a matrix with genetic distances
prova4
```

### Step 3: Mantel test

```
library(vegan)

mantel(xdis=DISTANCIES, ydis=prova4, method="pearson", permutations=999)
```

```
Results for G.australes Mantel test:

Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = DISTANCIES, ydis = prova4, method = "pearson",      permutations = 999)

Mantel statistic r: -0.00427
      Significance: 0.488

Upper quantiles of permutations (null model):
  90%   95% 97.5%   99%
0.125 0.169 0.222 0.247
Permutation: free
Number of permutations: 999
```

## Annex 3. Phylogenetic analysis

### 3.1 Selection of best evolution model with" phangorn" packages

```
model<-modelTest(dnaset)  # phangorn

aicmin<-min(model$AIC)
valuemin<-model[model$AIC==aicmin, ] #show the model with min AIC value
 valuemin
#resultado: el modelo evolutivo con el AIC más pequeño es el GTR+G
bicmin<-min(model$BIC)
modelbic<-model[model$BIC==bicmin, ] #show the model with min BIC value
modelbic
```

As a result, the best evolution model was GTR+R.

### 3.2 Estimation of genetic distances

```
dist.align<-dist.dna(dnaset) #not complicate models can be used, in that case
we use K80, which rate of mutation is the same for all nucleotides.

#save in excel file distances:
install.packages("xlsx") # paquete para crear excels.
install.packages("rJava")
library(xlsx)
library(rJava)
write.xlsx(distance.dna, distancias.xlsx)
```

### 3.3 Obtention of trees with ape and MEGA7.

• with distance methods NJ "ape":

```
# distance tree sin tener en cuenta los valores missing, secuencias cortas
treenj<-njs(dist.dna, model="K80")
class(treenj)
str(treenj)
plotnj<-plot(treenj, cex=0.2, sub="NJ tree") # plot the trees.
```

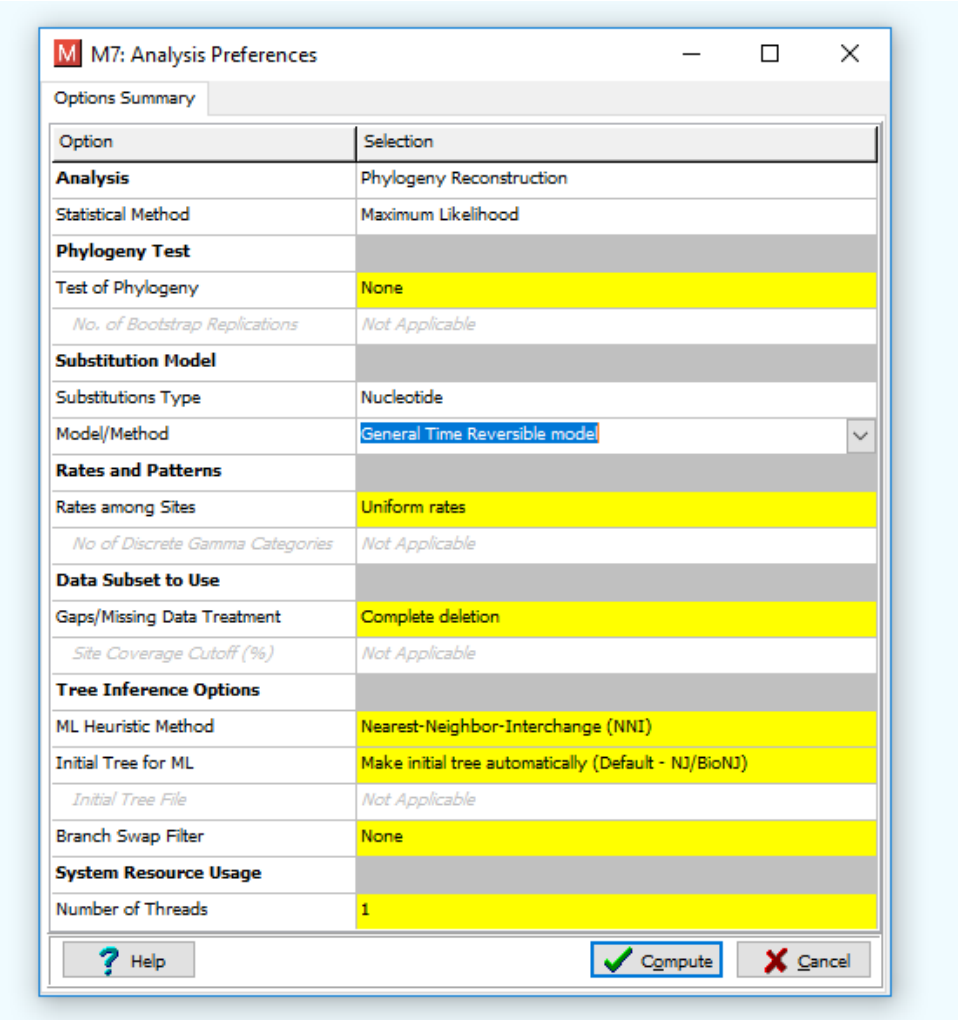
• with distance methods NJ "phangorn":

```
dnaphy<-as.phyDat(dnaset) # change the format to object phy, to pally
"phangorn" functions
distphy<-dist.ml(alignphy) #pairwise distances
temp <- as.data.frame(as.matrix(distphy))  #ceated numeric matrix to save in
excel file.
```

• with distance and trees with UPGMA methods "ape":

```
treeupgma<-upgma(distphy)
class(reeupgma)
plot.phylo(treeupgma, cex=0.2, sub= "UPGMA tree")
writeNexus(treeupgma, "treeupgma.nex") #we can create a nexus file, and open
with other programs with R is not very easy to plot.
```

Maximum Likelihood trees were elabotate with MEGA7, with evolution model GTR+G y and the option complete delection.



## Annex 4. Environmental data from GIS

Benthic.Max_Depth.Chlorophyll.Lt.max
Benthic.Max_Depth.Chlorophyll.Lt.min
Benthic.Max_Depth.Chlorophyll.Max
Benthic.Max_Depth.Chlorophyll.Mean
Benthic.Max_Depth.Chlorophyll.Min
Benthic.Max_Depth.Chlorophyll.Range
Benthic.Max_Depth.Current.Velocity.Lt.max
Benthic.Max_Depth.Current.Velocity.Lt.min
Benthic.Max_Depth.Current.Velocity.Max
Benthic.Max_Depth.Current.Velocity.Mean
Benthic.Max_Depth.Current.Velocity.Min
Benthic.Max_Depth.Current.Velocity.Range
Benthic.Max_Depth.Dissolved.oxygen.Lt.max
Benthic.Max_Depth.Dissolved.oxygen.Lt.min
Benthic.Max_Depth.Dissolved.oxygen.Max
Benthic.Max_Depth.Dissolved.oxygen.Mean
Benthic.Max_Depth.Dissolved.oxygen.Min
Benthic.Max_Depth.Dissolved.oxygen.Range
Benthic.Max_Depth.Iron.Lt.max
Benthic.Max_Depth.Iron.Lt.min
Benthic.Max_Depth.Iron.Max
Benthic.Max_Depth.Iron.Mean
Benthic.Max_Depth.Iron.Min
Benthic.Max_Depth.Iron.Range
Benthic.Max_Depth.Light.bottom.Lt.max
Benthic.Max_Depth.Light.bottom.Lt.min
Benthic.Max_Depth.Light.bottom.Max
Benthic.Max_Depth.Light.bottom.Mean
Benthic.Max_Depth.Light.bottom.Min
Benthic.Max_Depth.Light.bottom.Range
Benthic.Max_Depth.Nitrate.Lt.max
Benthic.Max_Depth.Nitrate.Lt.min
Benthic.Max_Depth.Nitrate.Max
Benthic.Max_Depth.Nitrate.Mean
Benthic.Max_Depth.Nitrate.Min
Benthic.Max_Depth.Nitrate.Range
Benthic.Max_Depth.Phosphate.Lt.max
Benthic.Max_Depth.Phosphate.Lt.min
Benthic.Max_Depth.Phosphate.Max
Benthic.Max_Depth.Phosphate.Mean
Benthic.Max_Depth.Phosphate.Min
Benthic.Max_Depth.Phosphate.Range
Benthic.Max_Depth.Phytoplankton.Lt.max
Benthic.Max_Depth.Phytoplankton.Lt.min
Benthic.Max_Depth.Phytoplankton.Max
Benthic.Max_Depth.Phytoplankton.Mean
Benthic.Max_Depth.Phytoplankton.Min
Benthic.Max_Depth.Phytoplankton.Range
Benthic.Max_Depth.Primary.productivity.Lt.max
Benthic.Max_Depth.Primary.productivity.Lt.min
Benthic.Max_Depth.Primary.productivity.Max
Benthic.Max_Depth.Primary.productivity.Mean
Benthic.Max_Depth.Primary.productivity.Min
Benthic.Max_Depth.Primary.productivity.Range
Benthic.Max_Depth.Salinity.Lt.max
Benthic.Max_Depth.Salinity.Lt.min
Benthic.Max_Depth.Salinity.Max
Benthic.Max_Depth.Salinity.Mean
Benthic.Max_Depth.Salinity.Min
Benthic.Max_Depth.Salinity.Range
Benthic.Max_Depth.Silicate.Lt.max
Benthic.Max_Depth.Silicate.Lt.min
Benthic.Max_Depth.Silicate.Max
Benthic.Max_Depth.Silicate.Mean
Benthic.Max_Depth.Silicate.Min
Benthic.Max_Depth.Silicate.Range
Benthic.Max_Depth.Temperature.Lt.max
Benthic.Max_Depth.Temperature.Lt.min
Benthic.Max_Depth.Temperature.Max
Benthic.Max_Depth.Temperature.Mean
Benthic.Max_Depth.Temperature.Min
Benthic.Max_Depth.Temperature.Range
Benthic.Mean_Depth.Chlorophyll.Lt.max
Benthic.Mean_Depth.Chlorophyll.Lt.min
Benthic.Mean_Depth.Chlorophyll.Max
Benthic.Mean_Depth.Chlorophyll.Mean
Benthic.Mean_Depth.Chlorophyll.Min
Benthic.Mean_Depth.Chlorophyll.Range
Benthic.Mean_Depth.Current.Velocity.Lt.max
Benthic.Mean_Depth.Current.Velocity.Lt.min
Benthic.Mean_Depth.Current.Velocity.Max
Benthic.Mean_Depth.Current.Velocity.Mean
Benthic.Mean_Depth.Current.Velocity.Min
Benthic.Mean_Depth.Current.Velocity.Range
Benthic.Mean_Depth.Dissolved.oxygen.Lt.max
Benthic.Mean_Depth.Dissolved.oxygen.Lt.min
Benthic.Mean_Depth.Dissolved.oxygen.Max
Benthic.Mean_Depth.Dissolved.oxygen.Mean
Benthic.Mean_Depth.Dissolved.oxygen.Min
Benthic.Mean_Depth.Dissolved.oxygen.Range
Benthic.Mean_Depth.Iron.Lt.max
Benthic.Mean_Depth.Iron.Lt.min
Benthic.Mean_Depth.Iron.Max
Benthic.Mean_Depth.Iron.Mean
Benthic.Mean_Depth.Iron.Min
Benthic.Mean_Depth.Iron.Range
Benthic.Mean_Depth.Light.bottom.Lt.max
Benthic.Mean_Depth.Light.bottom.Lt.min
Benthic.Mean_Depth.Light.bottom.Max
Benthic.Mean_Depth.Light.bottom.Mean
Benthic.Mean_Depth.Light.bottom.Min
Benthic.Mean_Depth.Light.bottom.Range
Benthic.Mean_Depth.Nitrate.Lt.max
Benthic.Mean_Depth.Nitrate.Lt.min
Benthic.Mean_Depth.Nitrate.Max
Benthic.Mean_Depth.Nitrate.Mean
Benthic.Mean_Depth.Nitrate.Min
Benthic.Mean_Depth.Nitrate.Range
Benthic.Mean_Depth.Phosphate.Lt.max
Benthic.Mean_Depth.Phosphate.Lt.min
Benthic.Mean_Depth.Phosphate.Max
Benthic.Mean_Depth.Phosphate.Mean
Benthic.Mean_Depth.Phosphate.Min
Benthic.Mean_Depth.Phosphate.Range
Benthic.Mean_Depth.Phytoplankton.Lt.max
Benthic.Mean_Depth.Phytoplankton.Lt.min
Benthic.Mean_Depth.Phytoplankton.Max
Benthic.Mean_Depth.Phytoplankton.Mean
Benthic.Mean_Depth.Phytoplankton.Min
Benthic.Mean_Depth.Phytoplankton.Range
Benthic.Mean_Depth.Primary.productivity.Lt.max
Benthic.Mean_Depth.Primary.productivity.Lt.min
Benthic.Mean_Depth.Primary.productivity.Max
Benthic.Mean_Depth.Primary.productivity.Mean
Benthic.Mean_Depth.Primary.productivity.Min
Benthic.Mean_Depth.Primary.productivity.Range
Benthic.Mean_Depth.Salinity.Lt.max
Benthic.Mean_Depth.Salinity.Lt.min
Benthic.Mean_Depth.Salinity.Max
Benthic.Mean_Depth.Salinity.Mean
Benthic.Mean_Depth.Salinity.Min
Benthic.Mean_Depth.Salinity.Range
Benthic.Mean_Depth.Silicate.Lt.max
Benthic.Mean_Depth.Silicate.Lt.min
Benthic.Mean_Depth.Silicate.Max

Benthic.Mean_Depth.Silicate.Mean
Benthic.Mean_Depth.Silicate.Min
Benthic.Mean_Depth.Silicate.Range
Benthic.Mean_Depth.Temperature.Lt.max
Benthic.Mean_Depth.Temperature.Lt.min
Benthic.Mean_Depth.Temperature.Max
Benthic.Mean_Depth.Temperature.Mean
Benthic.Mean_Depth.Temperature.Min
Benthic.Mean_Depth.Temperature.Range
Benthic.Min_Depth.Chlorophyll.Lt.max
Benthic.Min_Depth.Chlorophyll.Lt.min
Benthic.Min_Depth.Chlorophyll.Max
Benthic.Min_Depth.Chlorophyll.Mean
Benthic.Min_Depth.Chlorophyll.Min
Benthic.Min_Depth.Chlorophyll.Range
Benthic.Min_Depth.Current.Velocity.Lt.max
Benthic.Min_Depth.Current.Velocity.Lt.min
Benthic.Min_Depth.Current.Velocity.Max
Benthic.Min_Depth.Current.Velocity.Mean
Benthic.Min_Depth.Current.Velocity.Min
Benthic.Min_Depth.Current.Velocity.Range
Benthic.Min_Depth.Dissolved.oxygen.Lt.max
Benthic.Min_Depth.Dissolved.oxygen.Lt.min
Benthic.Min_Depth.Dissolved.oxygen.Max
Benthic.Min_Depth.Dissolved.oxygen.Mean
Benthic.Min_Depth.Dissolved.oxygen.Min
Benthic.Min_Depth.Dissolved.oxygen.Range
Benthic.Min_Depth.Iron.Lt.max
Benthic.Min_Depth.Iron.Lt.min
Benthic.Min_Depth.Iron.Max
Benthic.Min_Depth.Iron.Mean
Benthic.Min_Depth.Iron.Min
Benthic.Min_Depth.Iron.Range
Benthic.Min_Depth.Light.bottom.Lt.max
Benthic.Min_Depth.Light.bottom.Lt.min
Benthic.Min_Depth.Light.bottom.Max
Benthic.Min_Depth.Light.bottom.Mean
Benthic.Min_Depth.Light.bottom.Min
Benthic.Min_Depth.Light.bottom.Range
Benthic.Min_Depth.Nitrate.Lt.max
Benthic.Min_Depth.Nitrate.Lt.min
Benthic.Min_Depth.Nitrate.Max
Benthic.Min_Depth.Nitrate.Mean
Benthic.Min_Depth.Nitrate.Min
Benthic.Min_Depth.Nitrate.Range
Benthic.Min_Depth.Phosphate.Lt.max
Benthic.Min_Depth.Phosphate.Lt.min
Benthic.Min_Depth.Phosphate.Max
Benthic.Min_Depth.Phosphate.Mean
Benthic.Min_Depth.Phosphate.Min
Benthic.Min_Depth.Phosphate.Range
Benthic.Min_Depth.Phytoplankton.Lt.max
Benthic.Min_Depth.Phytoplankton.Lt.min
Benthic.Min_Depth.Phytoplankton.Max
Benthic.Min_Depth.Phytoplankton.Mean
Benthic.Min_Depth.Phytoplankton.Min
Benthic.Min_Depth.Phytoplankton.Range
Benthic.Min_Depth.Primary.productivity.Lt.max
Benthic.Min_Depth.Primary.productivity.Lt.min
Benthic.Min_Depth.Primary.productivity.Max
Benthic.Min_Depth.Primary.productivity.Mean
Benthic.Min_Depth.Primary.productivity.Min
Benthic.Min_Depth.Primary.productivity.Range
Benthic.Min_Depth.Salinity.Lt.max
Benthic.Min_Depth.Salinity.Lt.min
Benthic.Min_Depth.Salinity.Max
Benthic.Min_Depth.Salinity.Mean
Benthic.Min_Depth.Salinity.Min
Benthic.Min_Depth.Salinity.Range
Benthic.Min_Depth.Silicate.Lt.max
Benthic.Min_Depth.Silicate.Lt.min
Benthic.Min_Depth.Silicate.Max
Benthic.Min_Depth.Silicate.Mean
Benthic.Min_Depth.Silicate.Min
Benthic.Min_Depth.Silicate.Range
Benthic.Min_Depth.Temperature.Lt.max
Benthic.Min_Depth.Temperature.Lt.min
Benthic.Min_Depth.Temperature.Max
Benthic.Min_Depth.Temperature.Mean
Benthic.Min_Depth.Temperature.Min
Benthic.Min_Depth.Temperature.Range
Surface_Calcite.Mean
Surface_Chlorophyll.Lt.max
Surface_Chlorophyll.Lt.min
Surface_Chlorophyll.Max
Surface_Chlorophyll.Mean
Surface_Chlorophyll.Min
Surface_Chlorophyll.Range
Surface_Cloud.cover.Max
Surface_Cloud.cover.Mean
Surface_Cloud.cover.Min
Surface_Current.Velocity.Lt.max
Surface_Current.Velocity.Lt.min
Surface_Current.Velocity.Max
Surface_Current.Velocity.Mean
Surface_Current.Velocity.Min
Surface_Current.Velocity.Range
Surface_Diffuse.attenuation.Max
Surface_Diffuse.attenuation.Mean
Surface_Diffuse.attenuation.Min
Surface_Dissolved.oxygen.Lt.max
Surface_Dissolved.oxygen.Lt.min
Surface_Dissolved.oxygen.Max
Surface_Dissolved.oxygen.Mean
Surface_Dissolved.oxygen.Min
Surface_Dissolved.oxygen.Range
Surface_Ice.cover.Lt.max
Surface_Ice.cover.Lt.min
Surface_Ice.cover.Max
Surface_Ice.cover.Mean
Surface_Ice.cover.Min
Surface_Ice.cover.Range
Surface_Ice.thickness.Lt.max
Surface_Ice.thickness.Lt.min
Surface_Ice.thickness.Max
Surface_Ice.thickness.Mean
Surface_Ice.thickness.Min
Surface_Ice.thickness.Range
Surface_Iron.Lt.max
Surface_Iron.Lt.min
Surface_Iron.Max
Surface_Iron.Mean
Surface_Iron.Min
Surface_Iron.Range
Surface_Nitrate.Lt.max
Surface_Nitrate.Lt.min
Surface_Nitrate.Max
Surface_Nitrate.Mean
Surface_Nitrate.Min
Surface_Nitrate.Range
Surface_Par.Max
Surface_Par.Mean
Surface_pH
Surface_Phosphate.Lt.max
Surface_Phosphate.Lt.min
Surface_Phosphate.Max
Surface_Phosphate.Mean
Surface_Phosphate.Min

| | | |
|---|---|---|
| Surface_Phosphate.Range | Surface_Salinity.Lt.max | Surface_Temperature.Lt.min |
| Surface_Phytoplankton.Lt.max | Surface_Salinity.Lt.min | Surface_Temperature.Max |
| Surface_Phytoplankton.Lt.min | Surface_Salinity.Max | Surface_Temperature.Mean |
| Surface_Phytoplankton.Max | Surface_Salinity.Mean | Surface_Temperature.Min |
| Surface_Phytoplankton.Mean | Surface_Salinity.Min | Surface_Temperature.Range |
| Surface_Phytoplankton.Min | Surface_Salinity.Range | ETOPO1_Mean.Bathymetry |
| Surface_Phytoplankton.Range | Surface_Silicate.Lt.max | ETOPO1_Point.Bathymetry |
| Surface_Primary.productivity.Lt.max | Surface_Silicate.Lt.min | GEBCO_Mean.Bathymetry |
| Surface_Primary.productivity.Lt.min | Surface_Silicate.Max | GEBCO_Point.Bathymetry |
| Surface_Primary.productivity.Max | Surface_Silicate.Mean | Coral.Presence_Distance |
| Surface_Primary.productivity.Mean | Surface_Silicate.Min | Coral.Presence_Mean |
| Surface_Primary.productivity.Min | Surface_Silicate.Range | Coral.Presence_Point |
| Surface_Primary.productivity.Range | Surface_Temperature.Lt.max | |

## Annex 5. Environmental data obtention

In each sampling point, that was possible to find coordinates, environmental data was downloaded by ArcGIS (ESRI 2011, CA. Environmental Systems Research Institute), from the database Bio-ORACLE v2.0 (http://www.bio-oracle.org/). Layers downloaded were: *Surface*, *Benthic - Benthic - Minimum depth, Maximum depth Benthic - Average depth, Coral Reefs 2010* and *Bathymetry*. As a result, 315 *rasters* of environmental data were obtained.

All *rasters* were save in a excel file with R software following the next instructions:

#load database from the ".dbs" files generated by ArcGIS.

```
GIS.POINTS <- read.dbf("C:/Users/atudo/Desktop/POINTS/POINTS.dbf", as.is =
FALSE)
NAMES <- names(GIS.POINTS)
```

#create a file with all rasters.

```
for(F in 1:length(FILES)){
```

#Load DBF File & Modify Names

```
DBF.FILE <- read.dbf(paste0(FILES[F], ".dbf"), as.is = FALSE)
#Match Files (Add GIS Info to both GIS.POINTS & FULL.DATA
GIS.POINTS[, ncol(GIS.POINTS) + 1] <- DBF.FILE[match(GIS.POINTS$Code,
DBF.FILE$Code), ncol(DBF.FILE)]
}
```

#Rename Database

```
names(GIS.POINTS) <- c(NAMES, FILES)
```

#Export to Excel

```
write.xlsx2(as.data.frame(GIS.POINTS), file =
"C:/Users/atudo/Desktop/DATABASE/GIS POINTS.xlsx", sheetName = "GIS DATA",
col.names = TRUE, row.names = FALSE, append = FALSE, showNA = FALSE)
```

## Annex 6. Results of PCA

PCA were calculated with prcomp() following these steps:

- PCA equations estimations

- Check the variability of components and the importance of each variable to the components.

```
pcadata6<-prcomp(environvar2)
summary(pcadata6) #not scaling the data
```

```
Importance of components:
                            PC1      PC2      PC3      PC4      PC5      PC6
Standard deviation     1222.7395 61.1817 24.4815 15.14788 11.28698 6.15540
Proportion of Variance    0.9968  0.0025  0.0004  0.00015  0.00008 0.00003
Cumulative Proportion     0.9968  0.9993  0.9997  0.99983  0.99991 0.99994
                           PC7     PC8     PC9    PC10  PC11  PC12  PC13
Standard deviation     5.39013 4.80632 3.74489 3.36551 2.279 2.079 1.476
Proportion of Variance 0.00002 0.00002 0.00001 0.00001 0.000 0.000 0.000
Cumulative Proportion  0.99996 0.99997 0.99998 0.99999 1.000 1.000 1.000
```

```
pcadata6$rotation
```

```
round(pcadata6$rot[,1],2) #coefficients from the first component
```

```
          BM Chlo  BMD Current.Veloc   BMD Dis.oxygen  BM Depth.Iron        BMD Light
             0.00               0.00             0.00           0.00             0.00
      BMD Nitrate     BeMD Pho Range  BDM.Phos Range  BMD Phyto BMD Primary.productivity
            -0.01               0.00             0.00           0.00             0.00
         BMD Sal        BMD Sil Range  BMD.Sil Range    BMD Temp           SM Cal
             0.00              -0.03             0.00           0.00             0.00
          SM Chl     S Chlo.Min Range S Current.Veloc.Max SM Current.Velocity SM Diff.at
             0.00               0.00             0.00           0.00             0.00
SM Dissol oxygen SM Dissol oxygen Range        SM Iron     SM Nitrate          SM Par
             0.00               0.00             0.00           0.00             0.00
          SM pH          SM Pho Range     SM Phyto  SM_Primary.productivity   SM Sal
             0.00               0.00             0.00           0.00             0.00
          SM Sil         SM Temp Range          BAT     Coral dist
             0.00               0.00             1.00           0.00
```

```
round(pcadata6$rot[,2],2) #coefficients from the second component
```

```
          BM Chlo  BMD Current.Veloc   BMD Dis.oxygen  BM Depth.Iron        BMD Light
             0.00               0.00             0.85           0.00             0.05
      BMD Nitrate     BeMD Pho Range  BDM.Phos Range  BMD Phyto BMD Primary.productivity
            -0.20              -0.01             0.00           0.01             0.00
         BMD Sal        BMD Sil Range  BMD.Sil Range    BMD Temp           SM Cal
             0.02              -0.28            -0.04           0.07             0.00
          SM Chl     S Chlo.Min Range S Current.Veloc.Max SM Current.Velocity SM Diff.at
             0.00               0.00             0.00           0.00             0.00
SM Dissol oxygen SM Dissol oxygen Range        SM Iron     SM Nitrate          SM Par
             0.20               0.28             0.00          -0.02            -0.01
          SM pH          SM Pho Range     SM Phyto  SM_Primary.productivity   SM Sal
             0.00               0.00             0.00           0.00             0.01
          SM Sil         SM Temp Range          BAT     Coral dist
             0.00              -0.06            -0.01           0.17
```

```
data6<-data1[,SELECT6]
pcadata6<-prcomp(data6,scale=T)
summary(pcadata6)
pcadata6$rotation
```

# Results of PCA

```
round(pcadata6$rot[,1],2) #coefficients from the first component
```

```
          BM Chlo  BMD Current.Veloc   BMD Dis.oxygen  BM Depth.Iron        BMD Light
             0.13               0.00             0.27           0.18             0.12
      BMD Nitrate     BeMD Pho Range  BDM.Phos Range  BMD Phyto BMD Primary.productivity
            -0.25              -0.26            -0.02           0.16             0.11
         BMD Sal        BMD Sil Range  BMD.Sil Range    BMD Temp           SM Cal
             0.21              -0.17            -0.14           0.18             0.05
          SM Chl     S Chlo.Min Range S Current.Veloc.Max SM Current.Velocity SM Diff.at
            -0.04              -0.05            -0.19          -0.19             0.11
SM Dissol oxygen SM Dissol oxygen Range        SM Iron     SM Nitrate          SM Par
             0.22               0.24             0.26          -0.22            -0.22
          SM pH          SM Pho Range     SM Phyto  SM_Primary.productivity   SM Sal
             0.00              -0.24            -0.08          -0.13             0.18
          SM Sil         SM Temp Range          BAT     Coral dist
             0.00              -0.22             0.11           0.18
```

```
round(pcadata6$rot[,2],2) #coefficients from the second component
```

```
          BM Chlo       BMD Current.Veloc          BMD Dis.oxygen          BM Depth.Iron                BMD Light
             0.26                    0.25                    0.01                    0.04                    0.12
      BMD Nitrate                BeMD Pho          BDM.Phos Range          BMD Phyto BMD Primary.productivity
            -0.14                   -0.11                    0.06                    0.26                    0.27
         BMD Sal                 BMD Sil           BMD.Sil Range               BMD Temp                  SM Cal
            -0.08                   -0.10                    0.05                    0.21                    0.10
          SM Chl              S Chlo.Min      S Current.Veloc.Max      SM Current.Velocity                SM Diff.at
             0.32                    0.29                    0.18                    0.16                    0.25
SM Dissol oxygen   SM Dissol oxygen Range                SM Iron                SM Nitrate                  SM Par
            -0.04                    0.07                   -0.04                    0.11                   -0.04
          SM pH                   SM Pho                SM Phyto SM_Primary.productivity                  SM Sal
             0.09                    0.10                    0.31                    0.29                   -0.18
          SM Sil                  SM Temp                     BAT              Coral dist
             0.10                    0.05                    0.09                   -0.14
```

## Annex 7. Modelling geographical distribution

After creation the binary variables, models were performed following the next steps:
Step1:
-Create a matrix with all data, there is indications of each object involved in our model.
Step. 2:
-Proceeding of modelling our data
Step1: Create a matrix with all data, where are present indications of each object involved in our model:

Example of logistic model for *G.belizeanus:*

**Step1:** Create all objects to model

```
install.packages("biomod2")
library(biomod2)
# read data
datamod<-read.csv("C:/Users/Angi/Desktop/TEMP/DATABASE/GIS POINTS presencia
absencia.csv", header=TRUE, sep=";", stringsAsFactors = FALSE)
head(datamod) #with presence and absence information

datamod # here, all environmental data are selected, down will be removed,
only data selected by previous analysis.
attach(datamod)
# vector species presence/absence
spname1<-as.numeric(datamod$G.belizeanus)
length(spname1)
# vector coordinates
coordenates <- datamod[,c("longitud","latitud")]
# define environmental variables
environvar<-datamod[,SELECT6] # select only environmental data from the data.frame

names(environvar)<-SELECT7 # vector with abreviate names of environmental data
# formatting a matrix with your information
logmodel <- BIOMOD_FormatingData(resp.var =spname1,
```

```
expl.var = environvar,
resp.xy = coordenates,
resp.name = "G.belizeanus")
```

**Step2:** proceeding of modelling our data, selecting different options depending on your model.

```
myBiomodOption2 <- BIOMOD_ModelingOptions() # options models by default.
model2 <- BIOMOD_Modeling(
logmodel2,
models = c('GLM'),
models.options = myBiomodOption2,
NbRunEval=3,
DataSplit=80,
Prevalence=0.5,
VarImport=3,
models.eval.meth = c('TSS','ROC'),
SaveObj = TRUE,
rescal.all.models = TRUE,
do.full.models = FALSE)
# importance of variables from the model
get_variables_importance(model2)
get_variables_importance(mod)
attributes(model2)
```

The evaluation of the model was done with get_evaluations(), that True Skill Statistic (TSS), Receiveroperating characteristic (ROC), sensitivity and specificity (Allouche et al. 2006).

```
## Models evaluation
modeleval2 <- get_evaluations(model2)
modeleval2
dimnames(modeleval2) # types of evaluations
scores of evaluations of TSS and ROC were showed as:
modeleval2["TSS","Testing.data","RF",,]
modeleval2["ROC","Testing.data",,,]
```

```
    Testing.data Cutoff Sensitivity Specificity
TSS            1    669         100         100
ROC            1    668         100         100

, , GLM, RUN2, AllData

    Testing.data Cutoff Sensitivity Specificity
TSS        0.740  139.0         100          74
ROC        0.927  143.5         100          74

, , GLM, RUN3, AllData

    Testing.data Cutoff Sensitivity Specificity
TSS        0.647  535.0      66.667          98
ROC        0.823  539.5      66.667          98
```

**Figure 11. Evaluation of logistic model for *G.belizeanus***

## Annex 8. Results of data modelling (runs and evaluation of logistic models)

```
Model=GLM ( quadratic with no interaction )
        Stepwise procedure using AIC criteria
        selected formula : G.belizeanus ~
I(Benthic.Mean_Depth.Temperature.Mean^2) +
I(Benthic.Mean_Depth.Chlorophyll.Mean^2) +
    Surface_Par.Mean + GEBCO_Mean.Bathymetry

Benthic.Mean_Depth.Chlorophyll.Mean          0.115
Benthic.Mean_Depth.Temperature.Mean          0.330
Surface_Par.Mean                             0.944
GEBCO_Mean.Bathymetry                        0.502


   Testing.data Cutoff Sensitivity Specificity
TSS         0.82    140         100          82
ROC         0.95    144         100          82
```

*G. pacificus*

```
G.pacificus ~ (Coral.Presence_Distance^2) + Benthic.Mean_Depth.Silicate.Range
+  Surface_pH  +  I(Surface_Calcite.Mean^2)  +  Benthic.Mean_Depth.Iron.Mean  +
I(Benthic.Mean_Depth.Light.bottom.Mean^2) + I(Surface_Chlorophyll.Min^2)

Importance of the subsequent variables:
                                               GLM
Benthic.Mean_Depth.Iron.Mean                 0.435
Benthic.Mean_Depth.Light.bottom.Mean         0.413
Surface_Calcite.Mean                         0.197
Surface_Chlorophyll.Min                      0.196
Surface_pH                                   0.562
Coral.Presence_Distance                      0.929


Testing.data Cutoff Sensitivity Specificity
TSS         0.633  296.0         100      63.265
ROC         0.765  296.5         100      63.265
```

*Gambierdiscus ribotype 1*

Best model was: Gambierdiscus.ribotype.1 ~ Surface_pH

```
                                               GLM
Surface_pH                                   0.953


   Testing.data Cutoff Sensitivity Specificity
TSS         0.98  468.0         100      98.039
ROC         0.99  472.5         100      98.039
```

*Gambierdiscus ribotype 2*

```
Gambierdiscus.ribotype.2 ~ Surface_Silicate.Mean +
I(Benthic.Mean_Depth.Iron.Mean^2) +
    I(Surface_Silicate.Mean^2) + Benthic.Mean_Depth.Light.bottom.Mean +
    I(Surface_Par.Mean^2)
                                               GLM
Benthic.Mean_Depth.Iron.Mean                 0.280
Benthic.Mean_Depth.Phosphate.Range           0.008
Surface_Silicate.Mean                        0.954


   Testing.data Cutoff Sensitivity Specificity
TSS         0.88  464.5         100          88
ROC         0.91  465.0         100          88
```

## *Gambierdiscus* ribotype 4

RUN2:
Model=GLM ( quadratic with no interaction )
Stepwise procedure using AIC criteria
selected formula : Gambierdiscus.type.4 ~ Surface_Phosphate.Mean + Coral.Presence_Distance

| | GLM |
|---|---|
| Benthic.Mean_Depth.Chlorophyll.Mean | 0.115 |
| Benthic.Mean_Depth.Temperature.Mean | 0.330 |
| Surface_Par.Mean | 0.944 |
| GEBCO_Mean.Bathymetry | 0.502 |

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.913 | 440 | 100 | 91.304 |
| ROC | 0.957 | 444 | 100 | 91.304 |

RUN3:
Model=GLM ( quadratic with no interaction )
Stepwise procedure using AIC criteria
selected formula : Gambierdiscus.type.4 ~ Surface_Phosphate.Mean + I(Surface_Dissolved.oxygen.Mean^2)

| | GLM |
|---|---|
| Surface_Dissolved.oxygen.Mean | 0.198 |
| Surface_Par.Mean | 1.000 |
| GEBCO_Mean.Bathymetry | 0.328 |

, , GLM, RUN3, AllData

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.913 | 440 | 100 | 91.304 |
| ROC | 0.957 | 444 | 100 | 91.304 |

## *Gambierdiscus* ribotype 5

selected formula : Gambierdiscus.type.5 ~ Surface_Phosphate.Mean + I(Surface_Par.Mean^2)

| | GLM |
|---|---|
| Surface_Par.Mean | 0.602 |
| Surface_Phosphate.Mean | 0.544 |

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.827 | 404.0 | 100 | 82.692 |
| ROC | 0.904 | 404.5 | 100 | 82.692 |

## *Gambierdiscus ribotype 6*

selected formula : Gambierdiscus.type.6...G.toxicus ~ Surface_Phosphate.Mean + I(Benthic.Mean_Depth.Silicate.Mean^2)

| | GLM |
|---|---|
| Benthic.Mean_Depth.Silicate.Mean | 0.086 |
| Surface_Phosphate.Mean | 0.963 |

| Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|
| TSS | 0.788 | 401 | 100 | 78.846 |
| ROC | 0.885 | 406 | 100 | 78.846 |

## *G.lapillus*

Run : G.lapillus_AllData

G.lapillus_AllData_RUN1

selected formula : G.lapillus ~ I(GEBCO_Mean.Bathymetry^2) + Surface_pH

G.lapillus_AllData_RUN2

selected formula : G.lapillus ~ Surface_Calcite.Mean

G.lapillus_AllData_RUN3
selected formula : G.lapillus ~ Surface_pH + I(Surface_Diffuse.attenuation.Mean^2)
.

**RUN1, AllData**

| | GLM |
|---|---|
| Surface_pH | 1.00 |
| GEBCO_Mean.Bathymetry | 0.65 |

**, , RUN2, AllData**

| | GLM |
|---|---|
| Surface_Calcite.Mean | 1 |

**, , RUN3, AllData**

| | GLM |
|---|---|
| Surface_Diffuse.attenuation.Mean | 0.772 |
| Surface_pH | 1.000 |

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.981 | 490 | 100 | 98.077 |
| ROC | 0.981 | 494 | 100 | 98.077 |

, , GLM, RUN3, AllData

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 1 | 491 | 100 | 100 |
| ROC | 1 | 496 | 100 | 100 |

*G. carpenteri*

selected formula : G.carpenteri ~ I(Benthic.Mean_Depth.Salinity.Mean^2) + I(Benthic.Mean_Depth.Silicate.Range^2) +
Surface_Par.Mean + Benthic.Mean_Depth.Salinity.Mean

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.745 | 337 | 100 | 74.51 |
| ROC | 0.863 | 338 | 100 | 74.51 |

, , GLM, RUN2, AllData

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.784 | 332 | 100 | 78.431 |
| ROC | 0.882 | 337 | 100 | 78.431 |

, , GLM, RUN3, AllData

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.745 | 337 | 100 | 74.51 |
| ROC | 0.863 | 338 | 100 | 74.51 |

| | GLM |
|---|---|
| Benthic.Mean_Depth.Salinity.Mean | 0.857 |
| Benthic.Mean_Depth.Silicate.Range | 0.272 |
| Surface_Par.Mean | 0.640 |

*G.caribeaus*

```
selected formula : G.caribeaus ~ I(Surface_Silicate.Mean^2) +
Surface_Dissolved.oxygen.Range +
    I(Surface_Current.Velocity.Mean^2) + Surface_Calcite.Mean +
    I(Surface_pH^2) + Surface_Current.Velocity.Mean + Surface_Silicate.Mean +
    I(Surface_Temperature.Mean^2) + Surface_Temperature.Mean +
    I(Coral.Presence_Distance^2) + Benthic.Mean_Depth.Temperature.Mean


GLM, RUN1, AllData

    Testing.data Cutoff Sensitivity Specificity
TSS        0.633    435         100      63.265
ROC        0.816    438         100      63.265



                                          GLM
Benthic.Mean_Depth.Temperature.Mean         0.025
Surface_Calcite.Mean                        0.069
Surface_Current.Velocity.Mean               0.657
Surface_Dissolved.oxygen.Range              0.371
Surface_pH                                  0.385
Surface_Silicate.Mean                       0.393
Surface_Temperature.Mean                    0.517
Coral.Presence_Distance                     0.442
```

*G.excentricus*

```
G.excentricus ~ Surface_Silicate.Mean
+I(Benthic.Mean_Depth.Primary.productivity Mean^2) +
    Benthic.Mean_Depth.Light.bottom.Mean + Surface_Primary.productivity.Mean +
    I(Surface_Temperature.Mean^2)

Testing.data Cutoff Sensitivity Specificity
TSS         0.78    409         100          78
ROC         0.89    413         100          78


                                            GLM
Benthic.Mean_Depth.Light.bottom.Mean        0.313
Benthic.Mean_Depth.Primary.productivity.Mean 0.095
Surface_Primary.productivity.Mean           0.261
Surface_Silicate.Mean                       0.634
Surface_Temperature.Mean                    0.226
```

*G.australes*

```
G.australes ~ I(Surface_Temperature.Mean^2) + Benthic.Mean_Depth.Silicate.Mean
+ I(GEBCO_Mean.Bathymetry^2) + GEBCO_Mean.Bathymetry + Surface_Phosphate.Mean
+ I(Surface_Calcite.Mean^2) +I(Benthic.Mean_Depth.Primary.productivity.Mean^2)
+ I(Benthic.Mean_Depth.Phosphate.Range^2) +
I(Benthic.Mean_Depth.Light.bottom.Mean^2) + ISurface_pH +
I(Benthic.Mean_Depth.Chlorophyll.Mean^2)


                                            GLM
Benthic.Mean_Depth.Chlorophyll.Mean         0.205
Benthic.Mean_Depth.Light.bottom.Mean        0.050
Benthic.Mean_Depth.Phosphate.Range          0.087
Benthic.Mean_Depth.Primary.productivity.Mean 0.120
Benthic.Mean_Depth.Silicate.Mean            0.287
Surface_Calcite.Mean                        0.162
Surface_pH                                  0.270
Surface_Phosphate.Mean                      0.380
Surface_Temperature.Mean                    0.983
GEBCO_Mean.Bathymetry                       0.490



    Testing.data Cutoff Sensitivity Specificity
```

```
TSS        0.821  328.0         100      82.051
ROC        0.926  374.5         100      87.179
```

*F. paulensis*

```
selected formula : F.paulensis ~ I(Surface_Temperature.Mean^2) +
Surface_Dissolved.oxygen.Mean +
    I(Benthic.Mean_Depth.Primary.productivity.Mean^2) +
Benthic.Mean_Depth.Primary.productivity.Mean
                                      GLM
Benthic.Mean_Depth.Primary.productivity.Mean 0.301
Benthic.Mean_Depth.Silicate.Mean             0.868
Surface_Calcite.Mean                         0.319
Surface_Temperature.Mean                     0.687
GEBCO_Mean.Bathymetry                        0.474


  Testing.data Cutoff Sensitivity Specificity
TSS         0.92    445         100          92
ROC         0.96    450         100          92
```

*Gambierdiscus* spp.

```
Gambierdiscus ~ Surface_Dissolved.oxygen.Range +
I(Surface_Dissolved.oxygen.Range^2) +
    Surface_Current.Velocity.Mean +
I(Benthic.Mean_Depth.Primary.productivity.Mean^2) +
    I(Surface_Temperature.Mean^2) + Surface_Phosphate.Mean +
    I(Benthic.Mean_Depth.Silicate.Range^2) + Surface_Calcite.Mean +
    I(Benthic.Mean_Depth.Light.bottom.Mean^2)


-=-=-=--=-=-=- Gambierdiscus_AllData_RUN2
        selected formula : Gambierdiscus ~
I(Benthic.Mean_Depth.Silicate.Mean^2) + I(Surface_Calcite.Mean^2) +
    Surface_pH + Benthic.Mean_Depth.Silicate.Mean + GEBCO_Mean.Bathymetry +
    I(Benthic.Mean_Depth.Phosphate.Range^2)

                selected formula : Gambierdiscus ~
I(Benthic.Mean_Depth.Silicate.Mean^2) + I(Surface_Calcite.Mean^2) +
    Surface_pH + Benthic.Mean_Depth.Silicate.Mean + I(Surface_pH^2) +
    Surface_Chlorophyll.Min + GEBCO_Mean.Bathymetry +
I(Benthic.Mean_Depth.Phosphate.Range^2) +
    Surface_Current.Velocity.Mean

                                            GLM
Benthic.Mean_Depth.Light.bottom.Mean       0.034
Benthic.Mean_Depth.Primary.productivity.Mean 0.102
Benthic.Mean_Depth.Silicate.Range          0.066
Surface_Calcite.Mean                       0.029
Surface_Current.Velocity.Mean              0.225
Surface_Dissolved.oxygen.Range             1.000
Surface_Phosphate.Mean                     0.202
Surface_Temperature.Mean                   0.082

    Testing.data Cutoff Sensitivity Specificity
TSS         0.82    557          82         100
ROC         0.91    561          82         100

, , GLM, RUN2, AllData

    Testing.data Cutoff Sensitivity Specificity
TSS        0.427    540          76      66.667
ROC        0.713    544          76      66.667

, , GLM, RUN3, AllData

    Testing.data Cutoff Sensitivity Specificity
TSS        0.507  524.0          84      66.667
ROC        0.753  527.5          84      66.667
```

*Fukuyoa* spp.

```
Fukuyoa ~ Surface_Dissolved.oxygen.Range + I(Surface_Dissolved.oxygen.Range^2)
+ I(Surface_Phytoplankton.Mean^2) + Surface_Phosphate.Mean +
    Surface_Calcite.Mean + I(Surface_Current.Velocity.Mean^2) +
    I(Benthic.Mean_Depth.Light.bottom.Mean^2) + Surface_Par.Mean +
    Surface_Diffuse.attenuation.Mean
```

| | GLM |
|---|---|
| Benthic.Mean_Depth.Light.bottom.Mean | 0.251 |
| Surface_Calcite.Mean | 0.452 |
| Surface_Current.Velocity.Mean | 0.644 |
| Surface_Diffuse.attenuation.Mean | 0.449 |
| Surface_Dissolved.oxygen.Range | 0.976 |
| Surface_Par.Mean | 0.654 |
| Surface_Phosphate.Mean | 0.684 |
| Surface_Phytoplankton.Mean | 0.583 |

| | Testing.data | Cutoff | Sensitivity | Specificity |
|---|---|---|---|---|
| TSS | 0.940 | 449.0 | 100 | 94 |
| ROC | 0.977 | 449.5 | 100 | 94 |