

Lab Book

Cientific Initiation - Coral Metagenomes

Letícia Costa Cavalcante

Pedro Meirelles

Institute of Biology - UFBA

2018

Sumário

1	May 2018	3
1.1	13	3
1.1.1	Learning L ^A T _E X	3
1.1.2	Math environment	4
1.1.3	15 - Short-term project proposal	5
2	Creation of data base of metagenomes and genomes	6
2.1	28	6
2.1.1	Bibliographic search for genomes	6
2.1.2	Janeiro 2019	10
2.2	28	10
2.2.1	Bibliographic search for metagenomes	10
2.2.2	Amostras do MG-RAST indicadas pelo professor e Miguel	13
2.2.3	Janeiro de 2019	13
2.3	Mapa das amostras	14
3	Download of metagenomes	15
3.1	Download of mg-rast files	15
3.2	Download of NCBI metagenomes	15
3.2.1	Amostras indicadas por Miguel na reuniao de 30 de novembro	16
4	Format Conversion of NCBI metagenomes	17
4.0.1	Taxonprofiling	19
5	Adaptacao dos identificadores	20
6	Quality filter	21
6.0.1	Taxonprofiling	21
7	Uniformity filter (size and N bases)	22
7.1	Command line	22
7.1.1	Utilizacao do Taxonprofiling	23
7.1.2	Janeiro de 2019	23
8	Profilling metagenomes	25
8.1	Mg-Rast metagenomes	25
8.2	Kraken-biom	25

8.3	Teste com o kraken no scratch	26
8.4	Profiling no Atlantico com a ajuda do Rilquer	28
8.4.1	MG RAST metagenomes	28
8.4.2	SRA metagenomes	29
8.4.3	Janeiro de 2019	30
8.5	Profiling no Scolymia	31
8.6	Analises e obtencao de figuras	34
8.7	Obtencao de figuras com o segundo profiling feito com a ajuda do Rilquer - 23/10/2018	39
8.8	Obtencao de figuras 30/10/2018	45
8.8.1	FILO	47
8.8.2	FAMILIA	51
8.9	Figuras feitas em dezembro de 2018	56
8.10	Fevereiro de 2019	68
8.10.1	Graficos de abundancia	68
8.10.2	Problemas no Atlantico	73
8.10.3	Ferramenta microbiome no R	74
9	Functional annotation of metagenomes	75
10	references	77
11	Softwares, instalacao e linhas	79
11.1	Profilling metagenomes	79
12	Fundamentos teóricos e escrita do artigo	80
12.1	março de 2019	80
13	Meetings	82
14	Escrita do artigo	83
14.0.1	30 de outubro 2018	83
14.1	30 de novembro	84

Capítulo 1

May 2018

1.1 13

1.1.1 Learning L^AT_EX

- Working folder: *path*

L^AT_EX is a high-quality typesetting system, available as free software, which allows to produce scientific or technical documents *latex-main*. I am using L^AT_EX to create a Bioinformatics Lab Book. To compile my Lab Book, I can use command lines ([pdflatex](#) and [bibtex](#)). Afterwards I can visualise the produced *.pdf* file with evince or another reader. Alternatively, I can use a Latex editor, such as TexWorks (<https://www.tug.org/texworks/>), which allows me to write the code and control the *pdf* file in the same environment (Figure 1.1).

To compile the *.tex* file in the command line:

```
$pdflatex lab-book  
$bibtex lab-book  
$pdflatex lab-book  
$pdflatex lab-book
```

To visualise the *.pdf*:

```
$evince lab-book.pdf &
```

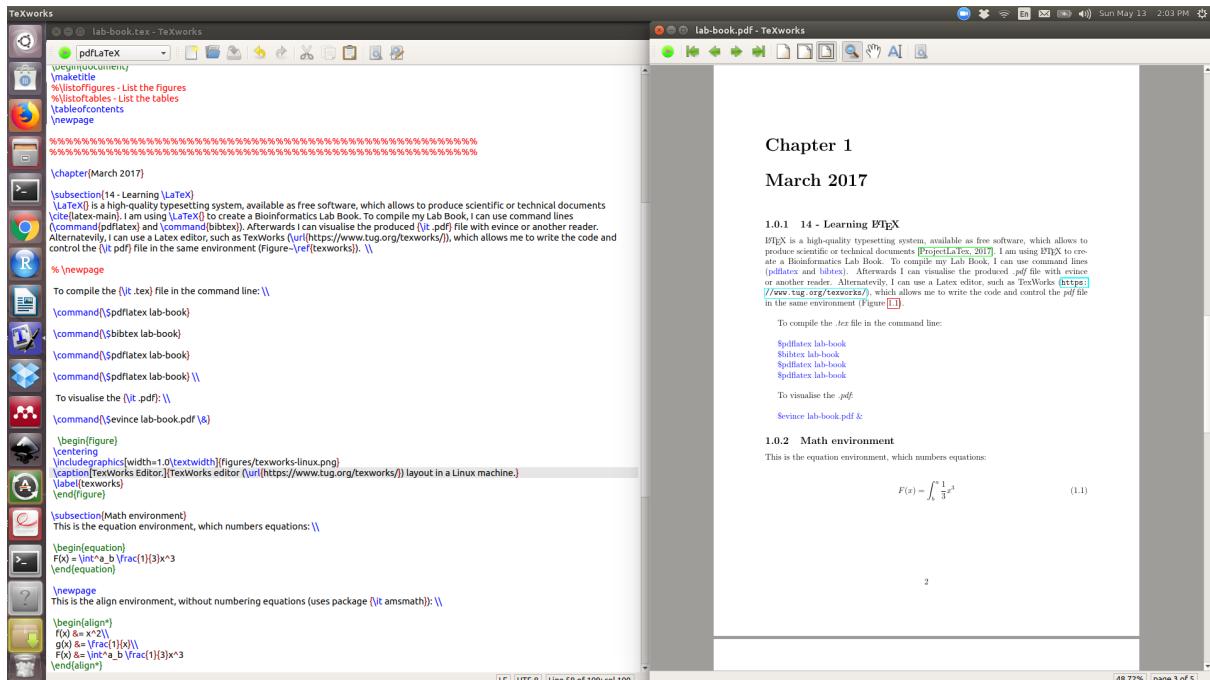


Figura 1.1: TexWorks editor (<https://www.tug.org/texworks/>) layout in a Linux machine.

1.1.2 Math environment

This is the equation environment, which numbers equations:

$$F(x) = \int_b^a \frac{1}{3}x^3 \quad (1.1)$$

This is the align environment, without numbering equations (uses package *amsmath*):

$$\begin{aligned}f(x) &= x^2 \\g(x) &= \frac{1}{x} \\F(x) &= \int_b^a \frac{1}{3}x^3\end{aligned}$$

1.1.3 15 - Short-term project proposal

Some text here. Including and referencing a table (table 1.1).

- First numbered list item
- Second numbered list item

Tabela 1.1: table0

species	changes	score
Macaque	4	0.0
Human	2	14.9
Orangutan	0	0.0
Pan	0	0.0
Gorilla	0	0.0

Capítulo 2

Creation of data base of metagenomes and genomes

2.1 28

2.1.1 Bibliographic search for genomes

Found a new possibility of phyla list. Because of this, there are four possibilities of list of microorganisms phyla, one of them, the SILVA database, is based in RNA sequences:

- The list of Prokariotic names with stading nomenclature <http://www.bacterio.net/-classifphyla.html>
- SILVA database LSU(large subunit of ribosome) <https://www.arb-silva.de/browser/lSU/>
- SILVA database SSU(small subunit of ribosome) <https://www.arb-silva.de/browser/ssu/>
- PATRIC GENOMES https://www.patricbrc.org/view/Taxonomy/2#view_tab=taxontree

The list of articles used until now is:

- 10.1038/nature14486
- 10.1038/ismej.2013.111
- 10.1038/ismej.2013.174
- 10.1038/ismej.2016.43
- 10.1038/nature12352
- 10.1038/nature14486
- 10.1038/nature21031
- 10.1038/ismej.2015.233

- 10.1038/ncomms13219
- 10.1073/pnas.0801980105
- 10.1111/1462-2920.13362
- 10.1126/science.1132690
- 10.1186/s40168-015-0077-6

The list os correspondent phyla and articles is above

Tabela 2.1: table 1

DOI	Phylum
10.1038/nature14486	Candidatus Falkowbacteria
10.1038/nature14486	Candidatus Kuenenbacteria
10.1038/nature14486	Candidatus Magasanikbacteria
10.1038/nature14486	Candidatus Uhrbacteria
10.1038/nature14486	Candidatus Moranbacteria
10.1038/nature14486	Candidatus Azambacteria
10.1038/nature14486	Candidatus Yanofskybacteria
10.1038/nature14486	Candidatus Jorgensenbacteria
10.1038/nature14486	Candidatus Wolfebacteria
10.1038/nature14486	Candidatus Giovannonibacteria
10.1038/nature14486	Candidatus Nomurabacteria
10.1038/nature14486	Candidatus Campbellbacteria
10.1038/nature14486	Candidatus Adlerbacteria
10.1038/nature14486	Candidatus Kaiserbacteria
10.1038/nature14486	C. S. yataiensis
10.1038/nature14486	Pacebacteria
10.1038/nature14486	Candidatus Collierbacteria
10.1038/nature14486	Candidatus Beckwithbacteria
10.1038/nature14486	Candidatus Roizmanbacteria
10.1038/nature14486	Candidatus Saphirobacteria
10.1038/nature14486	Candidatus Amesbacteria
10.1038/nature14486	Candidatus Woesebacteria
10.1038/nature14486	Candidatus Gottesmanbacteria
10.1038/nature14486	Candidatus Levybacteria
10.1038/nature14486	Candidatus Daviesbacteria
10.1038/nature14486	Candidatus Curtissbacteria
10.1038/nature14486	WWE3
10.1038/nature14486	CPR3
10.1038/nature14486	WS6
10.1038/nature14486	Candidatus Berkelbacteria
10.1038/nature14486	Candidatus Peregrinibacteria
10.1038/nature14486	Candidatus Gracilibacteria
10.1038/nature14486	CPR2
10.1038/nature14486	Kazan
10.1038/nature14486	Saccharibacteria (TM7)
10.1038/nature14486	SR1
10.1038/ncomms13219	Candidatus Kerfeldbacteria
10.1038/ncomms13219	Candidatus Komeilibacteria
10.1038/ncomms13219	Candidatus Andersenbacteria
10.1038/ncomms13219	Candidatus Ryanbacteria
10.1038/ncomms13219	Candidatus Niyogibacteria

10.1038/ncomms13219	Candidatus Tagabacteria
10.1038/ncomms13219	Candidatus Terrybacteria
10.1038/ncomms13219	Candidatus Vogelbacteria
10.1038/ncomms13219	Candidatus Zambryskibacteria
10.1038/ncomms13219	Candidatus Taylorbacteria
10.1038/ncomms13219	Candidatus Sungbacteria
10.1038/ncomms13219	Candidatus Brennerbacteria
10.1038/ncomms13219	Candidatus Spechtbacteria
10.1038/ncomms13219	Candidatus Staskawiczibacteria
10.1038/ncomms13219	Candidatus Wildermuthbacteria
10.1038/ncomms13219	Candidatus Portnoybacteria
10.1038/ncomms13219	Candidatus Woykebacteria
10.1038/ncomms13219	Candidatus Blackburnbacteria
10.1038/ncomms13219	Candidatus Chisholmbacteria
10.1038/ncomms13219	Candidatus Buchananbacteria
10.1038/ncomms13219	Candidatus Jacksonbacteria
10.1038/ncomms13219	Candidatus Veblenbacteria
10.1038/ncomms13219	Candidatus Nealsonbacteria
10.1038/ncomms13219	Candidatus Colwellbacteria
10.1038/ncomms13219	Candidatus Liptonbacteria
10.1038/ncomms13219	Candidatus Harrisonbacteria
10.1038/ncomms13219	Candidatus Yonathbacteria
10.1038/ncomms13219	Candidatus Lloydibacteria
10.1038/ncomms13219	Candidatus Abawacabacteria
10.1038/ncomms13219	Candidatus Doudnabacteria
10.1038/ismej.2013.111	Candidatus Poribacteria
10.1111/1462-2920.13362	Candidatus Desantisbacteria
10.1038/nature12352	Candidatus Omnitrophica
10.1038/nature12352	Candidatus Aminicenantes
10.1126/science.1132690	Candidatus Micrarchaeota
10.1038/nature14486	Candidatus Magasanikbacteria
10.1073/pnas.0801980105	Candidatus Korarchaeota
10.1038/nature12352	Candidatus Fervidibacteria
10.1038/nature12352	Candidatus Aenigmarchaeota
10.1038/ismej.2016.43	Candidatus Fermentibacteria
10.1038/ismej.2013.174	Candidatus Bathyarchaeota
10.1016/j.cub.2015.01.014	Candidatus Woesearchaeota
10.1016/j.cub.2015.01.014	Candidatus Kryptonia
10.1038/nature12352	Candidatus Diapherotrites
10.1038/nature12352	Candidatus Latescibacteria
10.1038/nature21031 10.1038/ismej.2015.233	Candidatus Thorarchaeota
10.1038/ncomms13219	Candidatus Lindowbacteria
10.1038/nature12352	Candidatus Parvarchaeota
10.1038/nature12352	Candidatus Cloacimonetes
10.1038/nature12352	Candidatus Hydrogenedentes
10.1038/nature12352	Candidatus Acetothermia

10.1038/nature12352	Candidatus Nanohaloarchaeota
10.1038/ncomms13219	Candidatus Eisenbacteria
10.1186/s40168-015-0077-6	candidate division WOR-3
10.1038/nature21031	Lokiarchaeota
10.1038/nature21031	Odinarchaeota
10.1038/nature21031	Heimdallarchaeota

2.1.2 Janeiro 2019

A quantidade de genomas aumentou razoavelmente, mas existem problemas na tabela e o Amaro pediu minha ajuda para melhorá-la. A primeira coisa da qual corri atrás foi das linhas que continham duas referencias de artigos. Fui no primeiro artigo: Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface e li um pouco por alto para encontrar a informação de construção. Acabei notando que o artigo tem o seguinte trecho: "The genome database was constructed from genomes, from metagenomes and from single-cell genomes (SAGs) collected in 2014. First, all curated, newly binned genomes from metagenomes (985 in total) were combined with 222 previously published genomes". Ou seja, 222 genomas não foram reconstruídos pelo artigo e são de uma referencia anterior, do artigo: "Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations" **Environ. Microbiol.** **19**, 459–474 (2017)." Indo no final desse artigo, encontrei referencias a dois bioprojects: PRJNA229517 and PRJNA297582. Ao ir nesse PRJNA229517, só encontrei listados Biosamples.

Como saber se o GenBank accession é correspondente ao identificador do Assembly Accession? R: Colocar o ID do Assembly accession para procurar em todas as bases de dados do ncbi e clicar em Assembly. Na página, virá o arquivo do GenBank e podemos checar o numero.

ARTZ00000000 - Essa sequencia foi removida e substituída por uma nova versao ARTZ00000000.1 - Esta destacada na planilha para falar para o Amaro, porque no artigo a sequencia esta com o identificador antigo.

2.2 28

2.2.1 Bibliographic search for metagenomes

The research for coral metagenomes started last year. The actual list is:

Tabela 2.2: table 1

IDs
mgm4440319.3
mgm4440370.3
mgm4440371.3

mgm4440372.3
mgm4440373.3
mgm4440374.3
mgm4440375.3
mgm4440376.3
mgm4440377.3
mgm4440378.3
mgm4440379.3
mgm4440380.3
mgm4440381.3
mgm4445755.3
mgm4445756.3
mgm4480739.3
mgm4480740.3
mgm4480741.3
mgm4480748.3
mgm4480750.3
mgm4487909.3
mgm4487910.3
mgm4487911.3
mgm4516541.3
mgm4516694.3
mgm4653307.3
mgm4694757.3
mgm4694758.3
mgm4694759.3
mgm4694760.3
SRR1275409
SRR1275449
SRR1283349
SRR1283371
SRR1283377
SRR1283433
SRR1283435
SRR1283437
SRR1286223
SRR1286225
SRR1286226
SRR1286227
SRR1286229
SRR1286232
SRR1822488
SRR1822516
SRR3499156
SRR3569370
SRR3694369

SRR3694370
SRR3694371
SRR3694372
SRR5215424
SRR5215454
SRR5215455
SRR5215456
SRR5215457
SRR5215458
SRR5215462
SRR5605611

I found these metagenomes in the article: "Metagenomic analysis reveals a green sulfur bacterium as a potential coral symbiont"

SRR2937345
SRR2937346
SRR2937347
SRR2937348
SRR2937349
SRR2937350
SRR2937351
SRR2937352
SRR2937353
SRR2937354
SRR2937355
SRR2937356

Espécie: *Platygyra carnosa* Healthy

I found other metagenomes of coral from article doi [10.3389/fmars.2018.00101](https://doi.org/10.3389/fmars.2018.00101) updated the file pmc_results_1.txt in the repository Lab_book. I continue to look the articles in results. Estou atualizando a lista pmc_results_2.txt Na pesquisa bibliografica olhando o título ja me faz perceber se devo descartar e olhar. E olho aqueles que marquei para olhar. Ao olhar, leio o resumo procurando por metodos. E vou para os metodos do artigo para checar. Checking the sizes of metagenomes files. The mg-rast metagenomes base have 72 Gb.

The pipeline of bioinformatic is different for MGRAST and NCBI. The size of NCBI should be superestimated, because the ncbi says the file size of sra file, but most of them is paired-end metagenomes, so when we apply fastq-dump, its generate two files fastq.

2.2.2 Amostras do MG-RAST indicadas pelo professor e Miguel

Na reunião feita em 30 de outubro, o professor e Miguel indicaram amostras existentes de Madracis para serem analisadas. O artigo é intitulado: "Turbulence-driven shifts in holobionts and planktonic microbial assemblages in St. Peter and St. Paul Archipelago, Mid-Atlantic Ridge, Brazil". DOI: 10.3389/fmicb.2015.01038. Amostras inclusas:

- mgm4486661.3
- mgm4486662.3
- mgm4486663.3
- mgm4486664.3
- mgm4486665.3
- mgm4486666.3
- mgm4486667.3
- mgm4486668.3
- mgm4486669.3

O estado de saúde dessas amostras não está claramente indicado no MG-RAST, mas pelo nome da amostra. No artigo, as amostras de Madracis branqueadas estão nomeadas com madble.

O professor indicou procurar pelo laboratório do professor Alexandre Rosado, para encontrar mais amostras de metagenomas de corais feitas no Brasil, especificamente nos trabalhos de Raquel Peixoto (não estava encontrando antes porque pensava que o nome era Raquel Rosado). Indo em seu perfil no google acadêmico, utilizei o recurso de ctrl f, pesquisando por coral. Os trabalhos que continham 'coral' no título não são de metagenoma.

2.2.3 Janeiro de 2019

Ao preparar os dados para refazer o mapa das amostras, notei coordenadas estranhas e notei problemas nas coordenadas na planilha de metadados. O primeiro que estou tentando solucionar são as coordenadas do trabalho de Dark Spot Syndrome. Eu copiando os dados da planilha para um arquivo novo e apaguei todos do trabalho de Dark Spot syndrome. Para recomeçar com esses dados do 0, eu baixei a BioInfoTable do estudos do SRA indicados pelo artigo "Corals and Their Microbiomes Are Differentially Affected by Exposure to Elevated Nutrients and a Natural Thermal Anomaly": SRP133535 for the 2014 metagenomes and SRP133699 for the 2012 metagenomes. A BioInfoTable de 2014 está nomeada como SraRunTable_2014.txt, a de 2012 SraRunTable_2012.txt. Comecei a editar a de 2014 e a primeira coisa que retirei foram linhas cujo número de bases e bytes eram 0, o que indica que eram metagenomas vazios. 2 metagenomas estavam nessas condições, então os retirei. Também retirei aqueles em que continham 'seawater metagenome' na coluna 'organism', pois interpretei que provavelmente eram metagenomas

de água. 6 metagenomas estavam assim. Coisas para segunda: terminar de preencher a nova tabela e separar duas partes em metadados ambientais e de bioinformatica, decidir a unidade armazenamento, gb ou mgb

2.3 Mapa das amostras

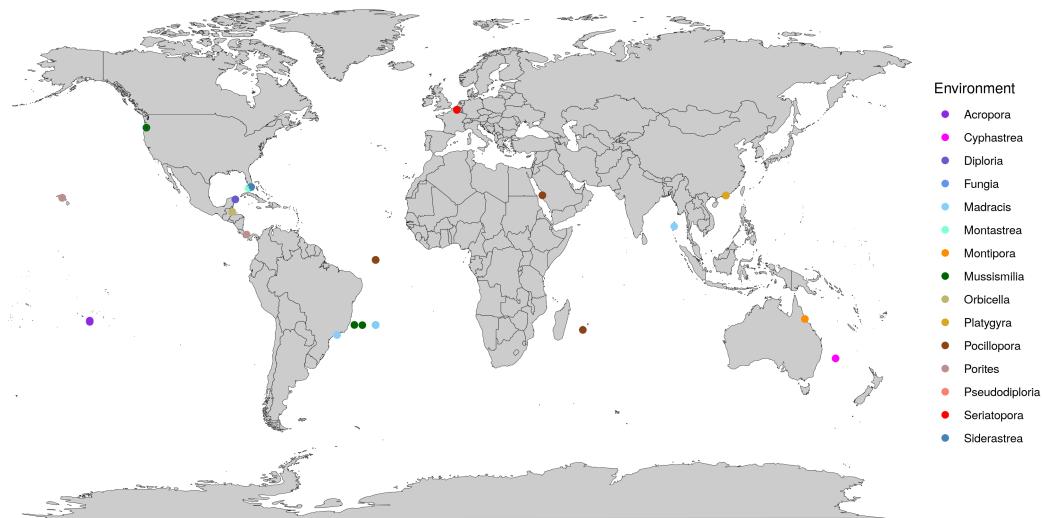


Figura 2.1: MAPA DOS METAGENOMAS

Capítulo 3

Download of metagenomes

3.1 Download of mg-rast files

Espaço no SDU Disponível para o ebiodiv: 10Tb Bia: 5Tb Rilquer: 2T Remanescente: 3Tb

- Working folder: *scratch/ebiodiv/leticia.cavalcante/mg_rast*

I insert the list of metagenomes in the files before using it. After this, I used the following command line:

- Command: *nohup bash download_curl_mgrast_corais.sh > download_curl_mgrast_corais.nohupout &*

3.2 Download of NCBI metagenomes

I use the script *download_sra_wget_corais.sh*, libs folder. I used the wget, because the curl is getting some problem in SDU. I noted that the size of the files is different:

Tabela 3.1: Comparing sizes of files

ID of metagenome	the size in NCBI site	size of file in SDU
SRR6785058	317.00 Mb	318M
SRR6785057	364.00 Mb	365M
SRR6785056	560.00 Mb	561M
SRR6785055	624.00 Mb	625M

So I checked the others files:

Tabela 3.2: Comparing sizes of files 2

ID of metagenome	the size in NCBI site	size of file in SDU	size of cleanned file
mgm4440319.3.299.1	30M	29.1 MB	28M
mgm4440370.3.299.1	3,6M	3.5 MB	3,5M
mgm4440371.3.299.1	5,0M	4.9 MB	4,8M
mgm4440372.3.299.1	6,0M	6.0 MB	5,9M
mgm4440373.3.299.1	6,2M	6.1 MB	6,0M
mgm4440374.3.299.1	4,1M	4.1 MB	4,0M
mgm4440375.3.299.1	3,8M	3.7 MB	3,7M
mgm4440376.3.299.1	3,9M	3.9 MB	3,8M
mgm4440377.3.299.1	3,5M	3.5 MB	3,4M
mgm4440378.3.299.1	6,2M	6.2 MB	6,1M
mgm4440379.3.299.1	7,0M	7.0 MB	6,9M
mgm4440380.3.299.1	5,2M	5.2 MB	5,2M
mgm4440381.3.299.1	6,4M	6.4 MB	6,4M
mgm4445755.3.299.1	158M	157.0 MB	155M
mgm4445756.3.299.1	150M	149.9 MB	147M
mgm4480739.3.299.1	8,0M	7.9 MB	7,9M
mgm4480740.3.299.1	12M	11.3 MB	12M
mgm4480741.3.299.1	8,5M	8.5 MB	8,5M
mgm4480742.3.299.1	10M	12.9 MB	10M
mgm4480743.3.299.1	15M	10.0 MB	14M
mgm4484839.3.299.1	13M	14.1 MB	13M
mgm4487909.3.299.1	17M	16.5 MB	17M
mgm4487910.3.299.1	36M	35.6 MB	36M
mgm4487911.3.299.1	12M	11.4 MB	12M
mgm4516541.3.299.1	161M	160.2 MB	163M
mgm4516694.3.299.1	193M	192.9 MB	193M
mgm4653307.3.299.1	17M	16.0 MB	17M
mgm4694757.3.299.1	1,9G	1.8 GB	1,9G
mgm4694758.3.299.1	2,2G	2.1 GB	2,2G
mgm4694759.3.299.1	1,7G	1.7 GB	1,8G
mgm4694760.3.299.1	592M	1.6 GB	597M

A Bia me informou que o SDU arredonda os valores de tamanho dos arquivos, entao ate o momento nao tive problemas com o download dos arquivos do mg_rast

3.2.1 Amostras indicadas por Miguel na reuniao de 30 de novembro

Fiz um loop para baixar as amostras chamado: loop_download_6_november.sh, funcionou no computador local, mas nao no servidor por algum motivo ligado a internet. Por isso, fiz o download das amostras no computador local e as transferi para o cluster Atlantico.

Capítulo 4

Format Conversion of NCBI metagenomes

Adaptei o script da Bia para fazer a conversao do dos arquivos .sra
Incialmente submeti apenas um na cpu_dev para testar:

```
Script: teste_slurm_job_fastq_dump_corais.sh
Numero do job: 220896
```

Deu certo.

O Rilquer me ajudou a criar um script que cria jobs de anotacao com o kraken2 para cada dois metagenomas.

O nome do script é 'creatijobfile.sh', ele unirá dois scripts: 'header' e 'ending'
Adaptei para criar jobs do fastq-dump para cada 2 arquivos .sra, haja visto que não tenho uma boa ideia do quanto cada fastq-dump demorará.

Jobs submetidos dia 27/09/2018:

- job_0.sh - job 221475
- job_100.sh - job 221476
- job_102.sh - job 221477
- job_104.sh - job 221478
- job_106.sh - job 221480
- job_108.sh - job 221481
- job_10.sh - job 221482
- job_110.sh - job 221483
- job_112.sh - job 221484
- job_114.sh - job 221485

- job_116.sh - job 221486
- job_118.sh - job 221487
- job_120.sh - job 221488
- job_122.sh - job 221489
- job_124.sh - job 221490
- job_126.sh - job 221493
- job_12.sh - job 221495

Jobs submetidos dia 09/10/2018:

- job_14.sh - slurm 226902
- job_16.sh - slurm 226903
- job_18.sh - slurm 226904
- job_20.sh - slurm 226905
- job_22.sh - slurm 226906
- job_24.sh - slurm 226907
- job_26.sh - slurm 226908
- job_28.sh - slurm 226909
- job_2.sh - slurm 226910
- job_30.sh - slurm 226911
- job_32.sh - slurm 226912

Jobs submetidos dia 19/10/2018:

- job_34.sh - slurm 231744
- job_36.sh - slurm 231745
- job_38.sh - slurm 231746
- job_40.sh - slurm 231747
- job_42.sh - slurm 231748
- job_44.sh - slurm 231749
- job_46.sh - slurm 231750
- job_48.sh - slurm 231751

- job_4.sh - slurm 231752
- job_50.sh - slurm 231753
- job_52.sh - slurm 231754
- job_54.sh - slurm 231755

4.0.1 Taxonprofiling

Essa subsecao foi escrita apos o Rilquer construir um script que unifica as etapas de bioinformatica em um script unico. Com script do Rilquer chamando Taxonprofiling, todas as etapas de bioinformatica estao unidas em um so script. Os jobs acimas estao guardados, mas com a utilizacao do script que o Rilquer construiu, as amostras estao sendo analisadas no Atlantico, cluster da UFBA

Capítulo 5

Adaptacao dos identificadores

Esse passo fez-se necessário nas análises anteriores, pois quando eu fazia a limpeza, os arquivos de outputs que saiam eram apenas os singletons. The output files were named as SRAXXX_good_singletons_1 and SRAXXX_good_singletons_2 and there aren't any other output files beyond these and the files with bad sequences. O Pablo fez a seguinte sugestão:

```
cat SRR1275409_pass_1.fastq | sed -r 's/(SRR1275409.[0-9]+)\.([0-9]+)/*\1_left/' > SRR1275409_pass_1.corr.fastq  
cat SRR1275409_pass_2.fastq | sed -r 's/(SRR1275409.[0-9]+)\.([0-9]+)/*\1_right/' > SRR1275409_pass_2.corr.fastq
```

Capítulo 6

Quality filter

This step is only required for NCBI metagenomes. The command line was proposed by Bia:

- trim_qual_left 25
- trim_qual_right 25

6.0.1 Taxonprofiling

Essa subsecao foi escrita apos o Rilquer construir um script que unifica as etapas de bioinformatica em um script unico. A etapa do quality filter foi modificada pelo Rilquer, apos ele insistir com o professor sobre qual seriam os melhores parametros de qualidade.

Segue criterios de qualidade:

- min_qual_score 13
- trim_qual_left 13
- trim_qual_right 13

Linha parcial abaixo:

```
-min_qual_score 13 -trim_qual_left 13 -trim_qual_right 13 -  
ns_max_n 2
```

Capítulo 7

Uniformity filter (size and N bases)

7.1 Command line

Parameters:

- min_len 80
- ns_max_p 2
- out_format 1

- Command: `nohup bash slurm_job_prinseq_single_cora_is_FASTA.bash &> slurm_prinseq_cora_is.out &`

Deu erro o job nohup: ignorando entrada

Location of PRINSEQ dir and scripts: /scratch/app/prinseq/0.20.4/bin srun Warning: can't run 1 processes on 21 nodes, setting nnodes to 1 srun Requested partition configuration not available now srun job 212425 queued and waiting for resources srun Force Terminated job 212425 srun Job has been cancelled srun error: Unable to allocate resources: No error srun Warning: can't run 1 processes on 21 nodes, setting nnodes to 1 srun Requested partition configuration not available now srun job 212428 queued and waiting for resources srun Force Terminated job 212428 srun Job has been cancelled

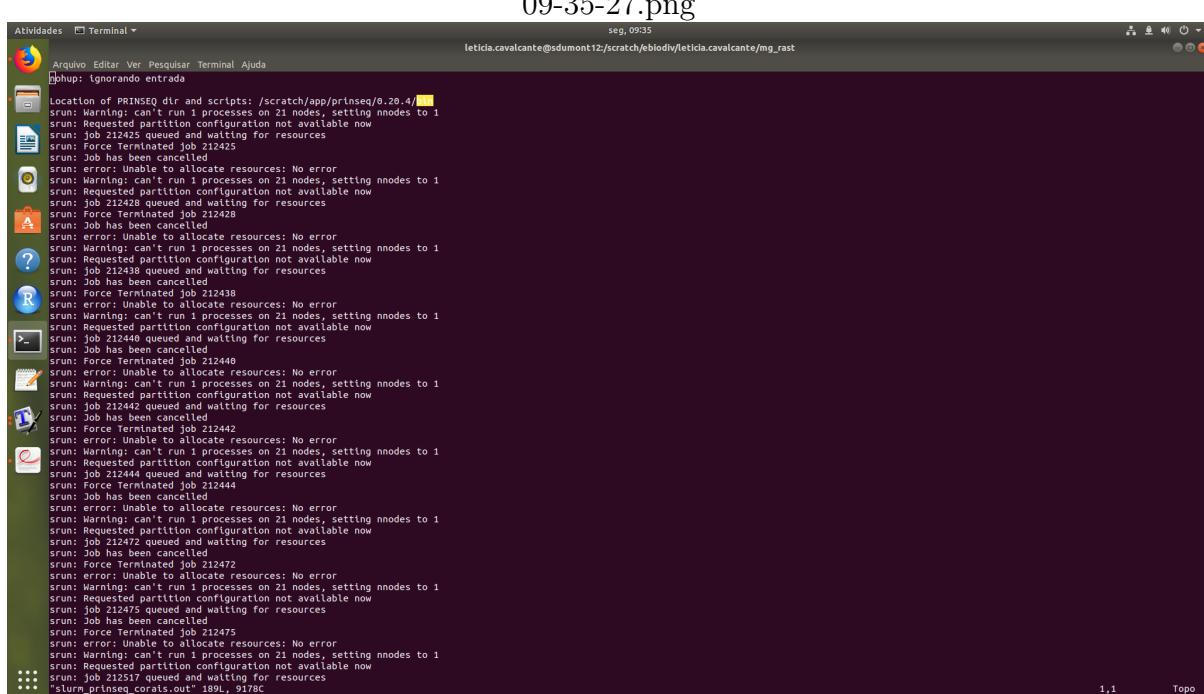


Figura 7.1: Erro no job no SDU

Ressubmeti o job com:

- Command: `sbatch slurm_job_prinseq_single_corais_FASTA.bash`

7.1.1 Utilizacao do Taxonprofiling

Essa subsecao foi escrita apos o Rilquer construir um script que unifica as etapas de bioinformatica em um script unico. Os criterios foram modificados para:

- `-min_len 50`
 - `-ns_max_n`

Linha total:

```
prinseq-lite -verbose -fastq ${OUTDIR}/fastqdump_output/${id}_pass_1.fastq -fastq2 ${OUTDIR}/fastqdump_output/${id}_pass_2.fastq -min_len 50 -min_qual_score 13 -trim_qual_left 13 -trim_qual_right 13 -ns_max_n 2 -out_format 1
```

7.1.2 Janeiro de 2019

Eu não encontrei alguns dos metagenomas limpos para fazer a re-anotacao, então eu preparei um script para poder fazer o prinseq isoladamente para as amostras do mg-rast. Existe uma copia local e no Atlantico. Para executar o prinseq, é necessario submeter um job em que haja a execucao do script que criei.

COPIA LOCAL

- Folder: /home/leticia/Documentos/libs/leticia_profiling_metagenomes
- Script: loop_prinseq_janeiro_16_2019.sh - Command: bash
loop_prinseq_janeiro_16_2019.sh

NO ATLANTICO

- Folder: /fsprofpedro/holobionts/mgrast/metagenomas
- Script: loop_prinseq_16_jan_2019.sh
- Script de submissao de job no Atlantico: job_prinseq_isolado_16_jan_2019.sh
- Numero de submissao: 153939.atlantico

Capítulo 8

Profilling metagenomes

8.1 Mg-Rast metagenomes

I used the following script in the following folder:

- Folder: *scratch/ebiodiv/leticia.ca valcante/mg_rast/filtered_prinseq_good*
- Command: *sbatch slurm_job_kraken2_cora is.sh*

The job doesn't work, o erro aparece na proxima figura

Ressubmeti o job, modificando a localizacao da DB do Kraken para a home do Rilquer.

Numero do job: 216410

Esse problema foi resolvido modificando o endereco da base para o scratch do Rilquer.

8.2 Kraken-biom

Pasta onde está instalado kraken-biom:

/home/leticia/.local/bin

Para executar: python2.7 .kraken-biom

Executar o help do kraken-biom:

kraken-biom -h

Abrir no vim o arquivo .bashrc e inserir:

```
export PATH=$PATH:/home/leticia/.local/bin/kraken-biom
```

Executar o help do kraken-biom:

kraken-biom -h

Eu fiz um teste da etapa "Creation of BIOM table of abundances" da pipeline da bia com os seguintes passos: Na pasta /home/leticia/Documentos/libs/leticia_profiling_metagenomes:

- kraken-biom selected_file -o table.biom -max D -min P
- biom convert -i table.biom -o table.from_biom_with_taxonomy.txt -to-tsv -header-key taxonomy

```

Atividades Terminal
Arquivo Editar Ver Pesquisar Terminal Ajuda
Terminal - leticia.cavalante@sdumont12:/scratch/ebiodiv/leticia.cavalante/mg_rast/filtered_prinseq_good
sex, 16:35

:rw----- 1 leticia.cavalante ebioldiv 1,8G Set 12 20:41 mgm4694759_prinseq_good_4Rbf.fasta
:rw----- 1 leticia.cavalante ebioldiv 597M Set 12 20:42 mgm4694760_prinseq_good_B0zy.fasta
:rw----r- 1 leticia.cavalante ebioldiv 556 Set 13 14:53 slurm-215145.out
:rw----r- 1 leticia.cavalante ebioldiv 9,6K Set 13 16:54 slurm-215343.out
:rw----r- 1 leticia.cavalante ebioldiv 3,4K Set 13 15:21 slurm_job_kraken2_corals.sh
[leticia.cavalante@sdumont12 filtered_prinseq_good]$ cat slurm-215343.out
sdumont1027 sdumont1028 sdumont1029 sdumont1066 sdumont1094 sdumont1206 sdumont1207 sdumont1276 sdumont1277 sdumont1312 sdumont1313 sdumont1314 sdumont1315 sdumont1316 sdumont1317 sdumont1318 sdumont1338 sdumont1339 sdumont1493 sdumont1494 sdumont5014 sdumont5015 sdumont5016 sdumont5017 sdumont5018 sdumont5019 sdumont5020 sdumont5021 sdumont5022 sdumont5023
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440319_prinseq_good_DYfg.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440370_prinseq_good_SlDp.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440371_prinseq_good_EDLk.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440372_prinseq_good_84wl.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440373_prinseq_good_JqUa.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440374_prinseq_good_9E3C.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440375_prinseq_good_Lvcn.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440376_prinseq_good_blxz.fasta!
srub: Warning: can't run 1 processes on 30 nodes, setting nnodes to 1
kraken2: database ("prj/ebioldiv/maria.costa/Kraken2_DB") does not contain necessary file taxo.k2d
srub: error: sdumont1027: task 0: Exited with exit code 2
Produced profile and report for file: mgm4440377_prinseq_good_02Ak.fasta!

```

Figura 8.1: 20 erro no job no SDU

- perl filterRank.pl input table.from_biom_with_taxonomy.txt --rank p > abundance.matrix

8.3 Teste com o kraken no scratch

Linha de teste:

perl selectGroups.pl input mgm4440370_prinseq_good_SiDP.fasta_kraken.report --file_groups groups.txt > selected_file

- First Command: *sbatch slurm_job_kraken2_corais.sh*
- Second Command:
kraken2 -db /prj/ebioldiv/rilquer.silva/Serrapilheira
/Kraken2_custom_DB/ mgm4440370_prinseq_good_SiDP.fasta
--output mgm4440370_prinseq_good_SiDP.fasta_kraken.profiled
--use-names --report mgm4440370_prinseq_good_SiDP.fasta_kraken.report

Já testei o comando acima na home do SDU e agora no scratch

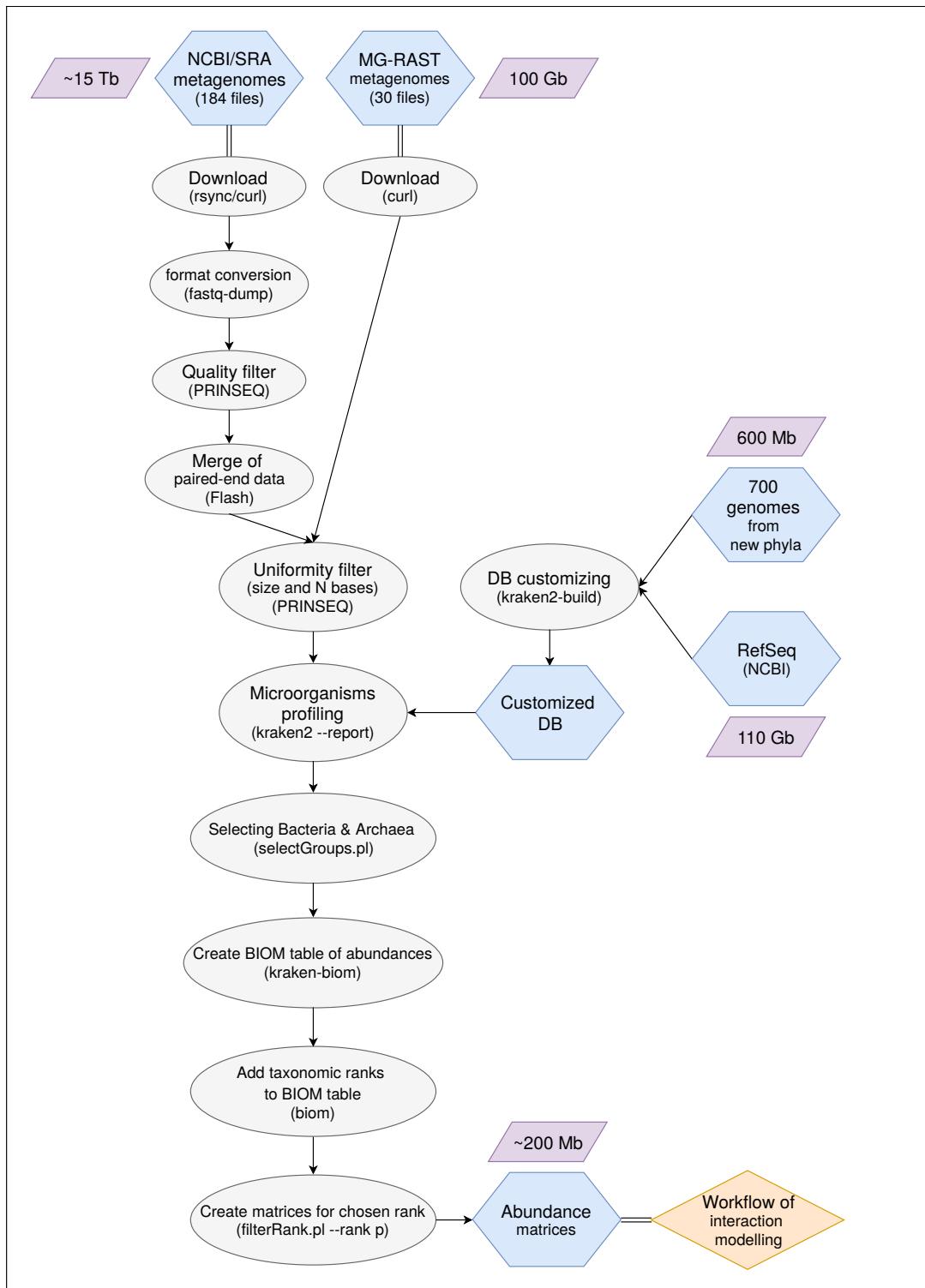


Figura 8.2: Pipeline of taxonomic annotation

8.4 Profiling no Atlantico com a ajuda do Rilquer

8.4.1 MG RAST metagenomes

O Rilquer fez um script que automatiza o processo, em que a limpeza e anotação ocorrem simultaneamente. Fiz um teste com esse script para um metagenoma com a seguinte linha:

- Comand: *taxonprofiling -s mgm4440378.3.299.1 -f MGRAST -k /home/pedro/-Kraken2_custom_DB*
- Script: taxonprofiling
- Folder: /fsprofpedro/holobionts/mgrast
- Para chamar o script: taxonprofiling

Sairam 4 outputs:

- mgm4440378_kraken_class
- mgm4440378_kraken_output
- mgm4440378_kraken_report
- mgm4440378_kraken_unclass

De acordo com a pipeline da Bia, o arquivo a ser usado é o report. Teste a seguir com a pasta com os metagenomas do mg-rast inteiro:

- Comand: *taxonprofiling -d /fsprofpedro/holobionts/mgrast -f MGRAST -k /home/-pedro/Kraken2_custom_DB*
- Folder: /fsprofpedro/holobionts

No email do Rilquer, vi que tenho que submeter o job para uma fila que não tem acesso ao fsprofpedro. Então criei uma pasta temporária chamada mgrast_temp, no home/pedro

- Job: *profiling_metagenomes_corais_mgrast.sh*
- Folder temporário em que as amostras foram copiadas: /home/pedro/mgrast_temp
- Folder de submissão: /home/pedro/

Na pasta /fsprofpedro/holobionts, tem um exemplo de job chamado jobexample. Submissão:

- Folder de submissão: /home/pedro/mgrast_temp
- Numero: 122296.atlantico
- Command: qsub profiling_metagenomes_corais_mgrast.sh

I have noticed that the files of abundance matrix is empty, I need to warn Rilquer. I had transferred the report files to this computer, to the following folder: /home/leticia/-Documentos/dados/report_atualizado_23_10_2018.

Sequence of command lines:

```
- Folder: /home/leticia/Documentos/libs/leticia_profiling_metagenomes  
Command 1: nohup bash select_groups_perl_corais.sh > select_groups_corais_perl.nohupout &  
Command 2: biom convert -i table_corais_mg_rast.biom -o table_corais_mg_rast.from_biom_with_taxonomy.txt -to-tsv -header-key taxonomy  
Command 3: perl filterRank.pl --input table_corais_mg_rast.from_biom_with_taxonomy.txt --rank p > abundance_corais_mgrast_2.matrix
```

Output gerado: **abundance_corais_mgrast_2.matrix**. O output gerado foi modificado no R, transposto, colocado em porcentagem e com os detalhes de genero, especie e estado de saude e retirei a linha gerada quando transformei em data frame, com X1, X2 e etc.

Coloquei para rodar novamente as amostras do MG-RAST para as matrizes de familia serem geradas. Job: 123190.atlantico

8.4.2 SRA metagenomes

Originalmente, os arquivos estão no folder: **/fsprofpedro/holobionts/SRA**. O total de arquivos é 158764768. Para fazer o trabalho de fastq dump, limpeza e anotacao, os arquivos devem estar no seguinte folder:

```
- Folder: /home/pedro/holobionts_temp/SRA  
- Command: qsub profiling_metagenomes_corais_SRA.sh  
- Linha: taxonprofiling -d /home/pedro/holobionts_temp/SRA -f SRA -t /home/pedro/sratoolkit.2.9.2-ubuntu64 -k /home/pedro/DB -o /home/pedro/holobionts_temp/SRA/ -r p,f
```

Submeti o job para amostras do SRA no Atlantico, job 123186.atlantico. Notei que o job gerou os arquivos limpos, mas não encontro as matrizes de abundancias. Pedi ajuda do Rilquer. Testei uma linha isoladamente.

```
taxonprofiling -s /home/pedro/holobionts_temp/SRA/SRR6793730.sra  
-f SRA -t /home/pedro/sratoolkit.2.9.2-ubuntu64 -k /home/pedro/DB -o /home/pedro/holobionts_temp/SRA/ -r p,c,o,f,g
```

[breaklines]

```

Atividades Terminal
Arquivo Editar Ver Pesquisar Terminal Ajuda
Good bases (singletons file 1): 376,852
Good mean length (singletons file 1): 301.00
Good sequences (singletons file 2): 5 (0.00%)
Good bases (singletons file 2): 1,505
Good mean length (singletons file 2): 301.00
Bad sequences (file 1): 875,252 (99.86%)
Bad bases (file 1): 263,450,852
Bad mean length (file 1): 301.00
Bad sequences (file 2): 876,499 (100.00%)
Bad bases (file 2): 263,826,199
Bad mean length (file 2): 301.00
Sequences filtered by specified parameters:
min_qual_score: 1751751
Prinseq successfully run on sample SRR6793730
/home/pedro/bin/taxonprofiling: line 438: [: !=: unary operator expected
/home/pedro/bin/taxonprofiling: line 438: [: !=: unary operator expected
Loading database information... done.
5 sequences (0.00 Mbp) processed in 0.114s (2.6 Kseq/m, 1.59 Mbp/m).
5 sequences classified (100.00%)
0 sequences unclassified (0.00%)
Kraken successfully run for sample SRR6793730
Traceback (most recent call last):
  File "/home/pedro/miniconda2/bin/kraken-biom", line 11, in <module>
    sys.exit(main())
  File "/home/pedro/miniconda2/lib/python2.7/site-packages/kraken_biom.py", line 377, in main
    biomT = create_biom_table(sample_counts, taxa)
  File "/home/pedro/miniconda2/lib/python2.7/site-packages/kraken_biom.py", line 196, in create_biom_table
    generated_by=gen_str, input_is_dense=True)
  File "/home/pedro/miniconda2/lib/python2.7/site-packages/biom/table.py", line 508, in __init__
    errcheck(self)
  File "/home/pedro/miniconda2/lib/python2.7/site-packages/biom/err.py", line 472, in errcheck
    raise ret
biom.exception.TableException: Duplicate sample IDs!
Usage: biom convert [OPTIONS]
Try "biom convert -h" for help.

Error: Invalid value for "-i" / "--input-fp": File "/home/pedro/holobionts_temp/SRA//SRR6793730_table.biom" does not exist.
It was not possible to open file /home/pedro/holobionts_temp/SRA//SRR6793730_table.biom.tsv

```

Figura 8.3: Erro com o script taxon profiling se manifestando ao não gerar as matrizes de abundância

8.4.3 Janeiro de 2019

O Rilquer me informou de problemas na base de dados, então o profiling deve ser refeito. Para facilitar o trabalho, perguntei a ele se ele queria que os metagenomas limpos estivessem em algum local específico para facilitar a re-anotação. Ele disse que sim. O folder de destino que ele pediu foi: /fsprofpedro/holobiontes_limpos

Para não ter de copiar tudo à mão, fiz um script para copiar os metagenomas. Existem duas cópias desse script, uma no computador local, no folder: /home/leticia/-Documentos/libs/leticia_profiling_metagenomes, o nome é loopinho.sh. No Atlântico, está no folder: /fsprofpedro/holobionts/SRA/output/fastqdump_output, nome: **loopinho_copiar.sh**.

- Script: loopinho_copiar.sh
- Command: nohup bash loopinho_copiar.sh
- Folder: /fsprofpedro/holobionts/SRA/output/fastqdump_output

8.5 Profiling no Scolymia

- Folder: */home/scolymia/leticia/fsprofpedro/
holobionts/SRA/output/fastqdump_output/limpos*

O Felipe começou a me ajudar com a anotação taxonômica no Scolymia. Entretanto, devido a menor poder de computação, devo anotar amostras por vez. A lista de metagenomas é:

SRR1275409
SRR1275449
SRR1283349
SRR1283371
SRR1283377
SRR1283433
SRR1283435
SRR1283437
SRR1286223
SRR1286225
SRR1286226
SRR1286227
SRR1286229
SRR1286232
SRR1822488
SRR1822516
SRR2937345
SRR2937346
SRR2937347
SRR2937348
SRR2937349
SRR2937350
SRR2937351
SRR2937352
SRR2937353
SRR2937354
SRR2937355
SRR2937356
SRR3499156
SRR3569370
SRR3694369
SRR3694370
SRR3694371
SRR3694372
SRR5215424
SRR5215455
SRR5215457

SRR5215462
SRR5605611
SRR6784973
SRR6784974
SRR6784975
SRR6784976
SRR6784977
SRR6784978
SRR6784979
SRR6784980
SRR6784981
SRR6784982
SRR6784983
SRR6784984
SRR6784985
SRR6784986
SRR6784987
SRR6784988
SRR6784989
SRR6784990
SRR6784991
SRR6784992
SRR6784993
SRR6784994
SRR6784995
SRR6784996
SRR6784997
SRR6784998
SRR6784999
SRR6785000
SRR6785005
SRR6785006
SRR6785009
SRR6785010
SRR6785011
SRR6785012
SRR6785013
SRR6785014
SRR6785015
SRR6785017
SRR6785018
SRR6785019
SRR6785020
SRR6785021
SRR6785022
SRR6785023

SRR6785024
SRR6785026
SRR6785027
SRR6785028
SRR6785029
SRR6785030
SRR6785031
SRR6785032
SRR6785033
SRR6785034
SRR6785035
SRR6785036
SRR6785037
SRR6785038
SRR6785039
SRR6785040
SRR6785041
SRR6785042
SRR6785043
SRR6785044
SRR6785045
SRR6785046
SRR6785047
SRR6785048
SRR6785049
SRR6785050
SRR6785051
SRR6785052
SRR6785053
SRR6785054
SRR6785055
SRR6785056
SRR6785057
SRR6785058

Dessa lista, eu fui até SRR2937348 no dia 07/03/2019. Eu coloquei para rodar o script 'kraken_isolado_fevereiro_sra_2019_20_primeiros.sh' e já criei os outros.

8.6 Analises e obtencao de figuras

Apliquei o tutorial do professor para obtenção de figuras no R para visualização dos resultados.

- Script: *analisys.R*
- Folder: */home/leticia/Documentos/libs/R*

Figuras obtidas:

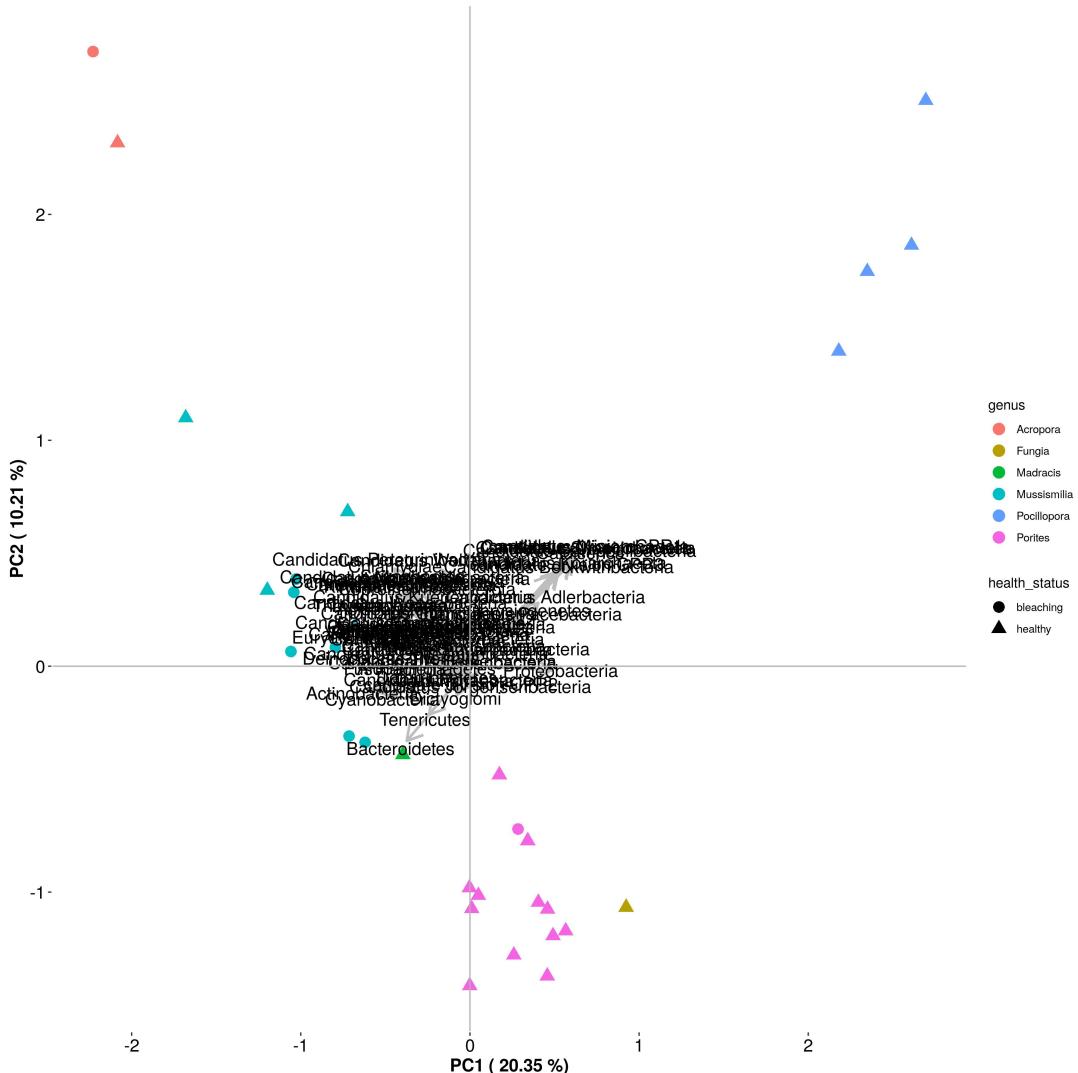


Figura 8.4: Análise de componentes principais com todos os filos como variaveis

Na analise acima, as variaveis são muitas e ficam muito sobrepostas, fazendo com que haja grande poluicao visual. O professor recomendou em marco a utilizar uma analise de Random Forest, para que as variaveis mais importantes para os metagenomas que trabalho sejam ranqueadas. O random forest é um algoritmo de machine learning que, a partir das duas categorias de saúde (categorias de supervisao), elencará as variáveis mais

importantes para classificar as amostras nesses dois estados. o random forest abaixo é supervisionado por estado de saúde

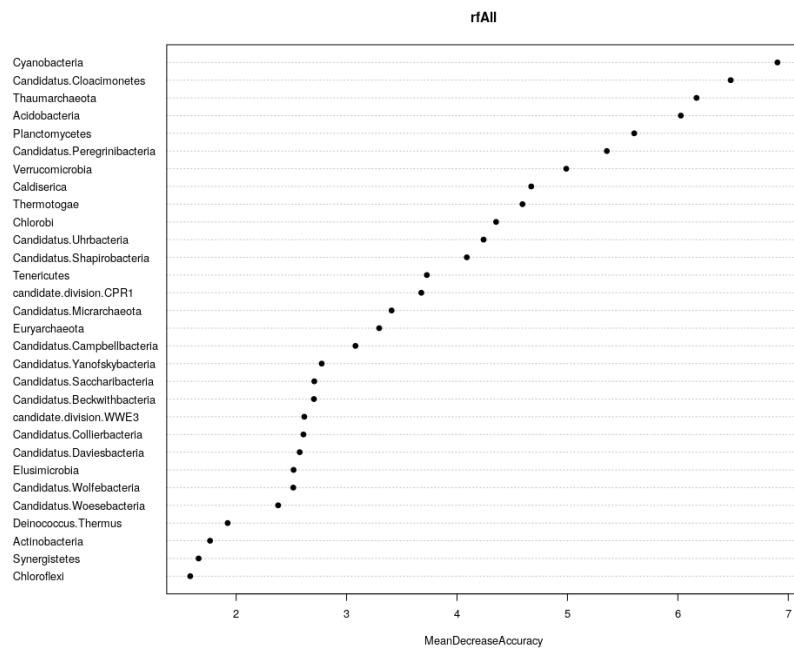


Figura 8.5: Random Forest ranqueando filos

Eu utilizei os 20 primeiros filos ranqueados para fazer o PCA. Segue esse PCA abaixo:

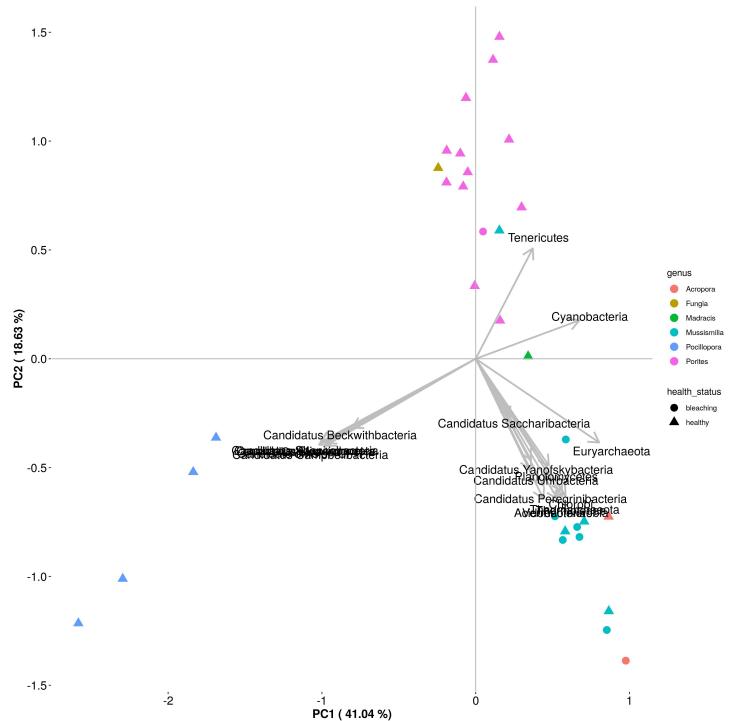


Figura 8.6: PCA com 20 filos utilizados no PCA

Uma tendencia se manteve: foi a separação das amostras por gênero. Mas a poluição visual ainda continuou, por isso fiz um random forest com os 15 primeiros filos indicados pelo random forest. Segue abaixo:

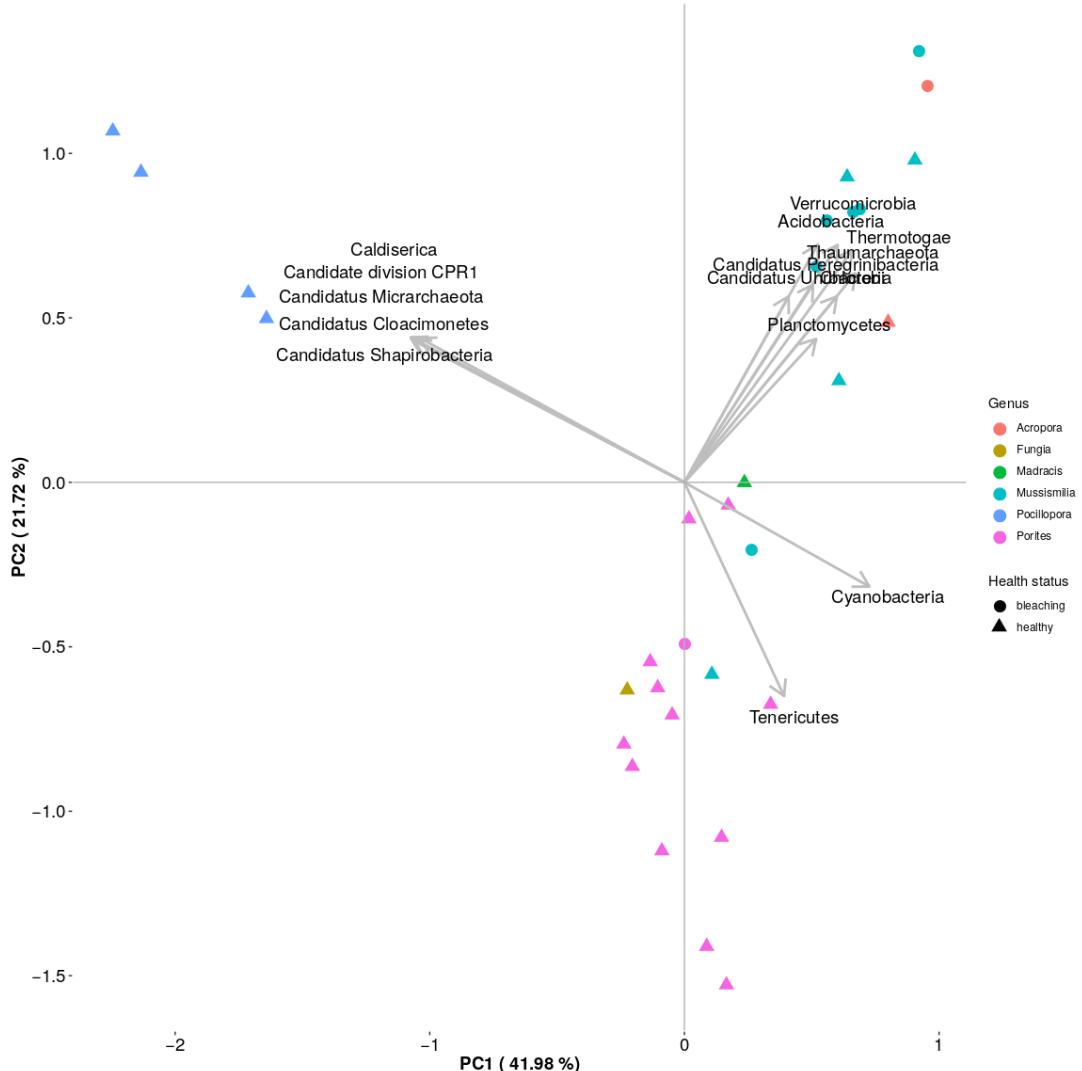


Figura 8.7: PCA com 15 filos como variáveis

Algumas tendências também se mantiveram e a explicação dos eixos melhorou levemente. As amostras agrupadas no quadrante direito superior são mais diferentes das que estão no quadrante esquerdo do que das que estão no quadrante inferior direito. Na reunião feita no dia 03/10/2018, o Amaro, o professor e Miguel me sinalizaram que existe uma separação forte entre gêneros, indicando que os grupos candidatos podem ser gênero - específicos. Surgiu a sugestão de leitura de textos em core microbiome e especificidade de filos entre gêneros e o professor sugeriu fazer um random forest não supervisionado que segue abaixo. O professor Garcia no congresso sugeriu utilizar os autovalores do PCA para ver quais podem ser mais relevantes (?).

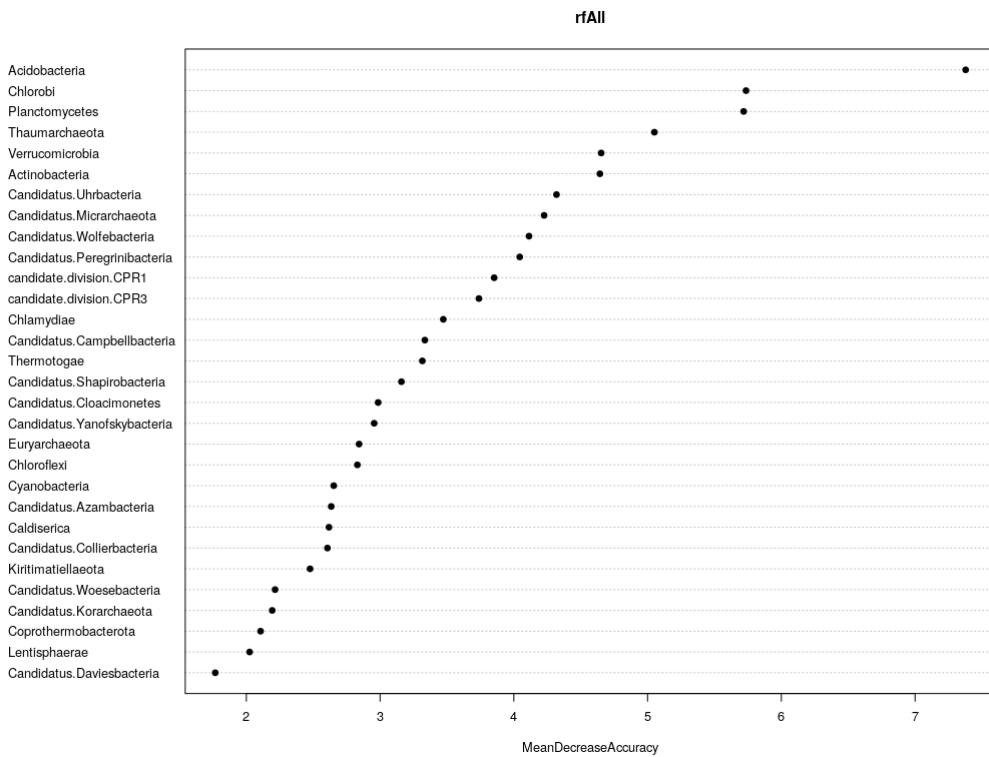


Figura 8.8: Random Forest nao supervisionado

Fiz um pca (libs/R/analisis.R) a partir dos primeiros 15 filos indicados no Random Forest nao supervisionado acima. Segue abaixo:

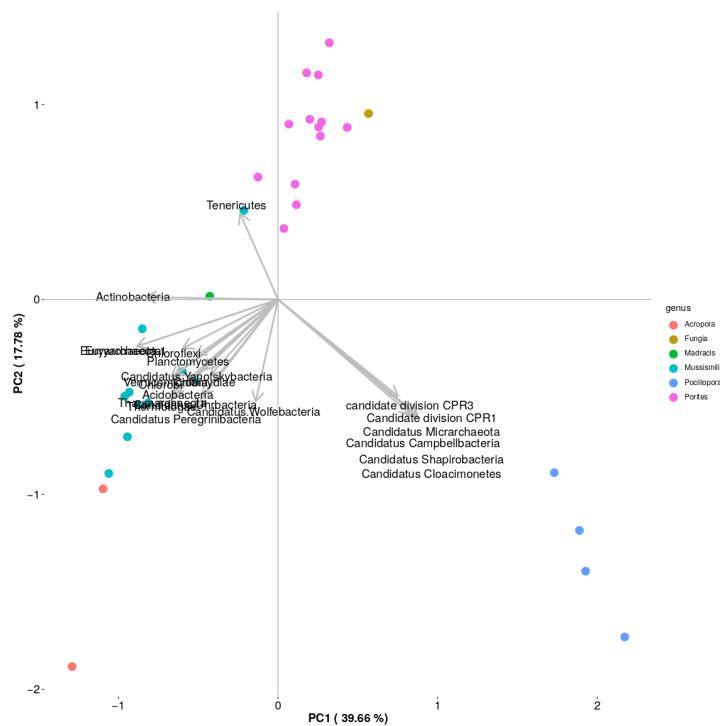


Figura 8.9: PCA a partir dos 20 filos primeiros filos que aparecem acima no random forest

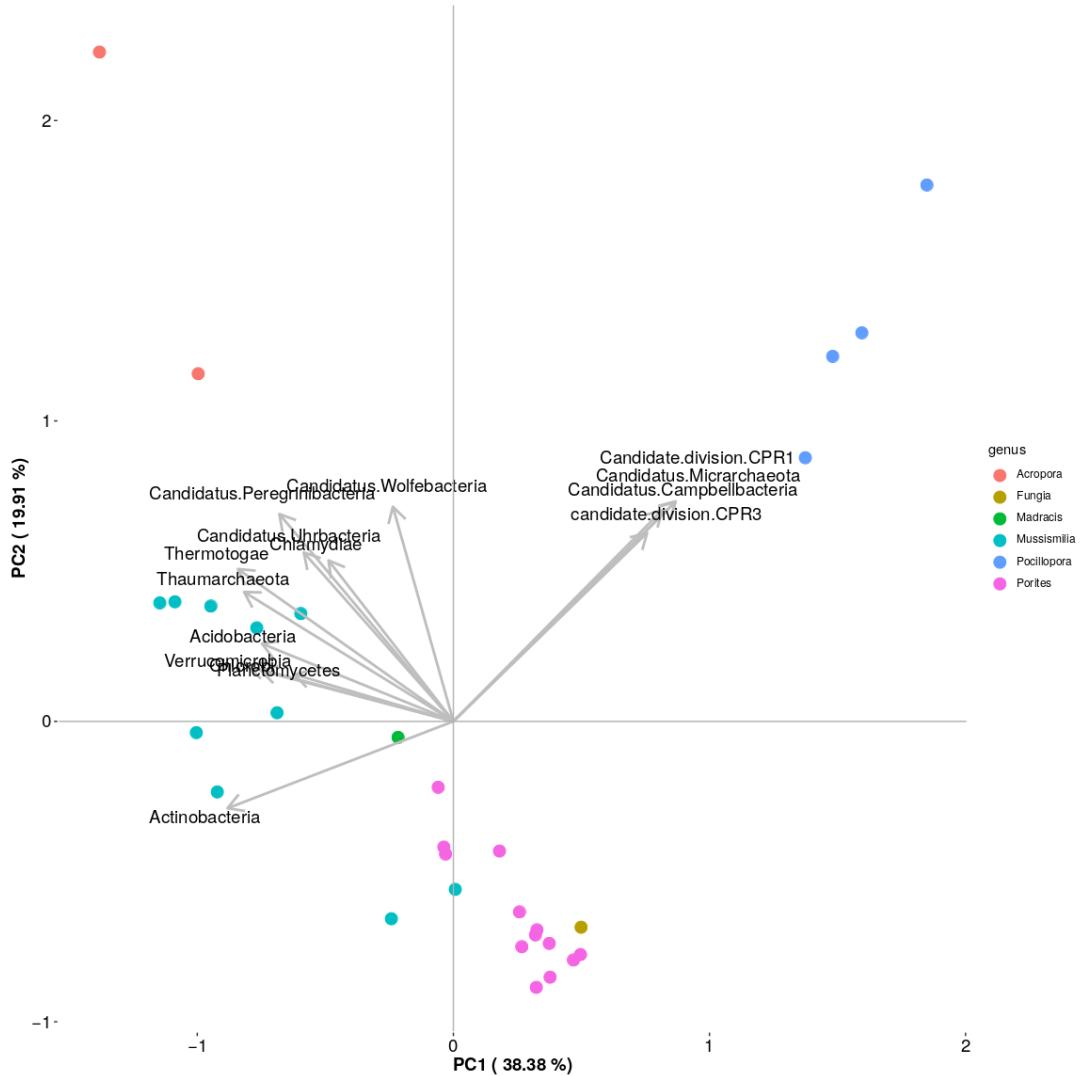


Figura 8.10: PCA a partir dos 15 filos primeiros filos que aparecem acima no random forest nao supervisionado

8.7 Obtencao de figuras com o segundo profiling feito com a ajuda do Rilquer - 23/10/2018

As figuras a seguir foram feitas com dados de abundancia gerados com a ajuda do Rilquer. As primeiras serao com abundancia dos metagenomas do MG-RAST.

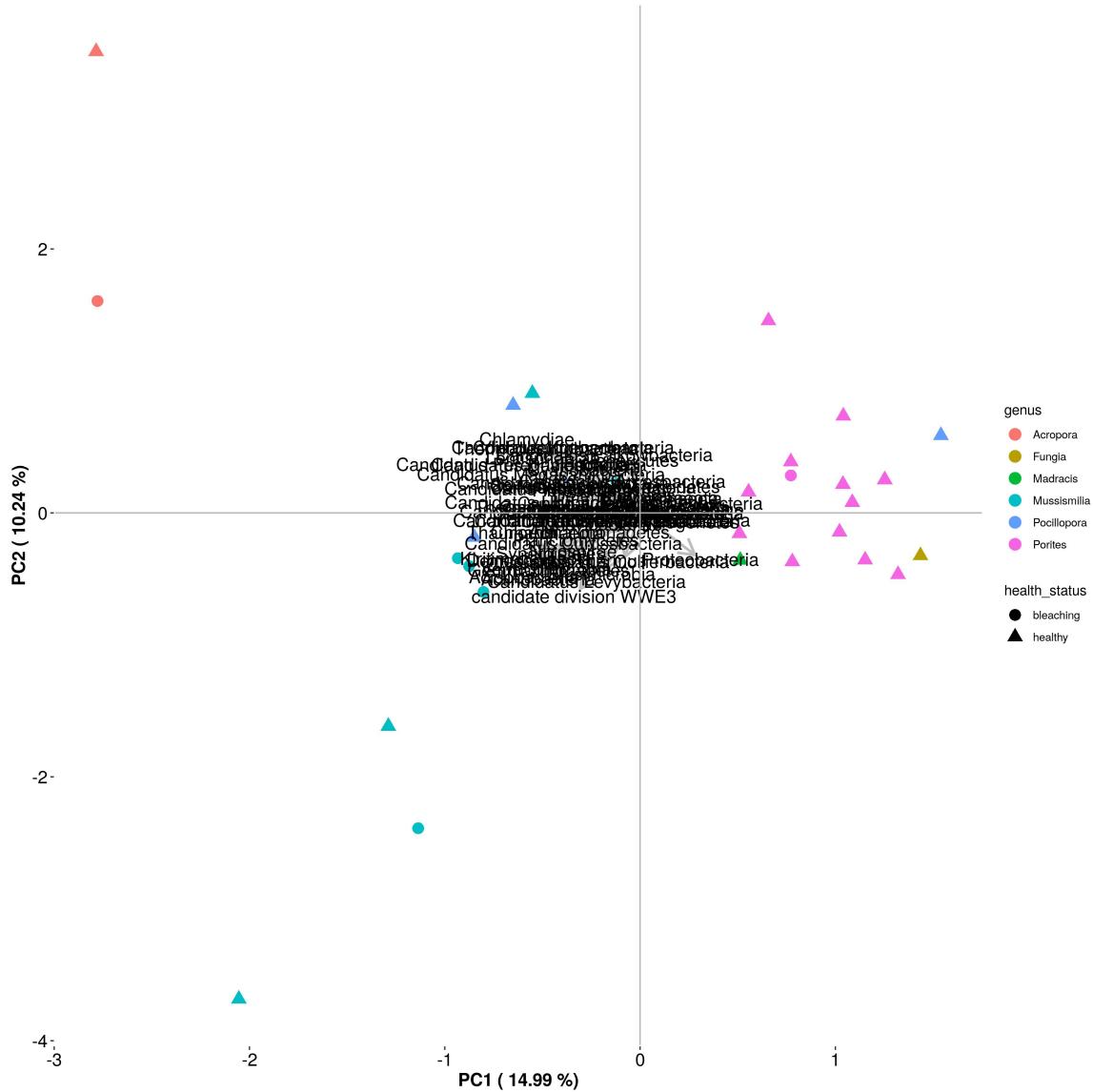


Figura 8.11: PCA das amostras de metagenomas do mg rast com todos os filos

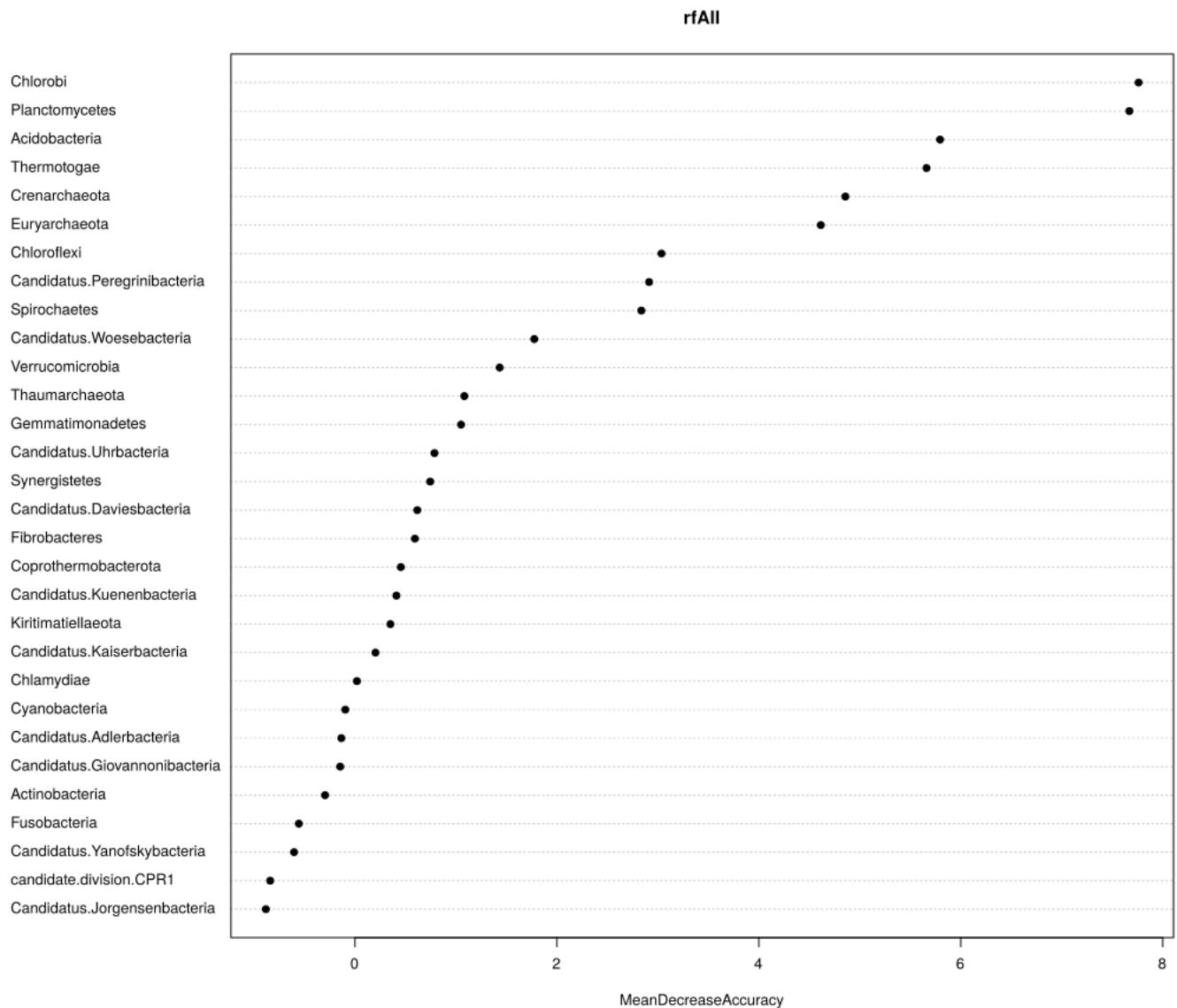


Figura 8.12: Random Forest não supervisionado dos metagenomas de corais do MG-RAST analisadas com a base definitiva

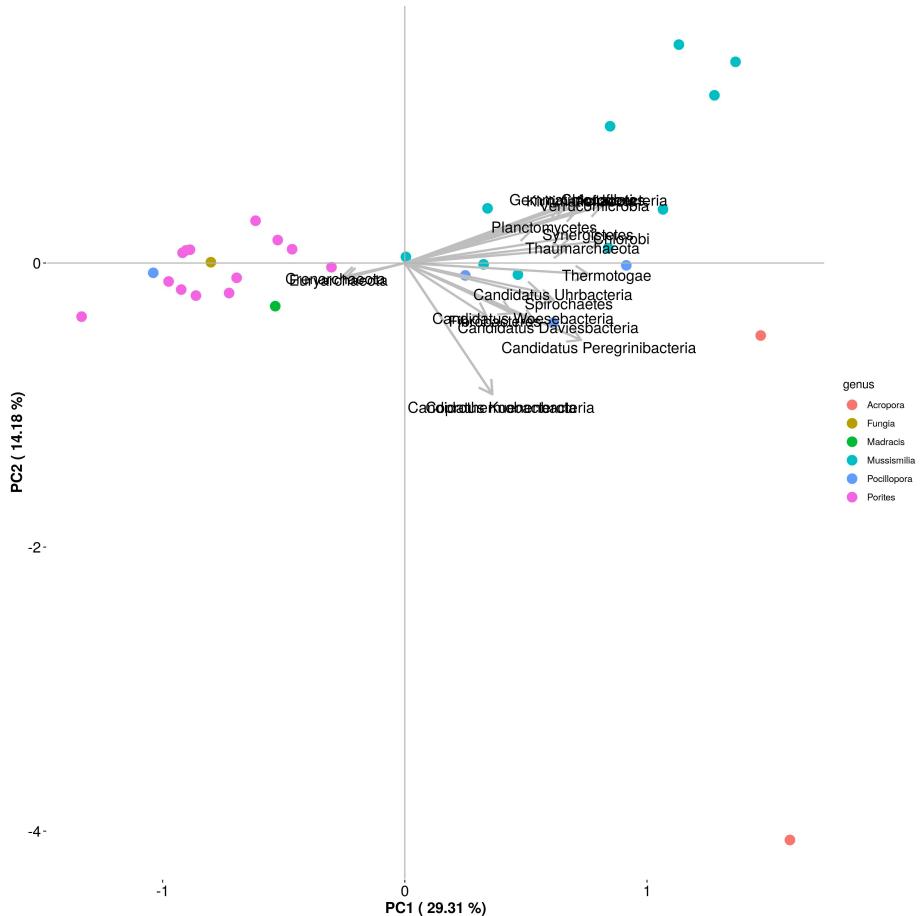


Figura 8.13: PCA GERADO COM OS PRIMEIROS 20 FILOS DO RANDOM FOREST NAO SUPERVISIONADO ACIMA

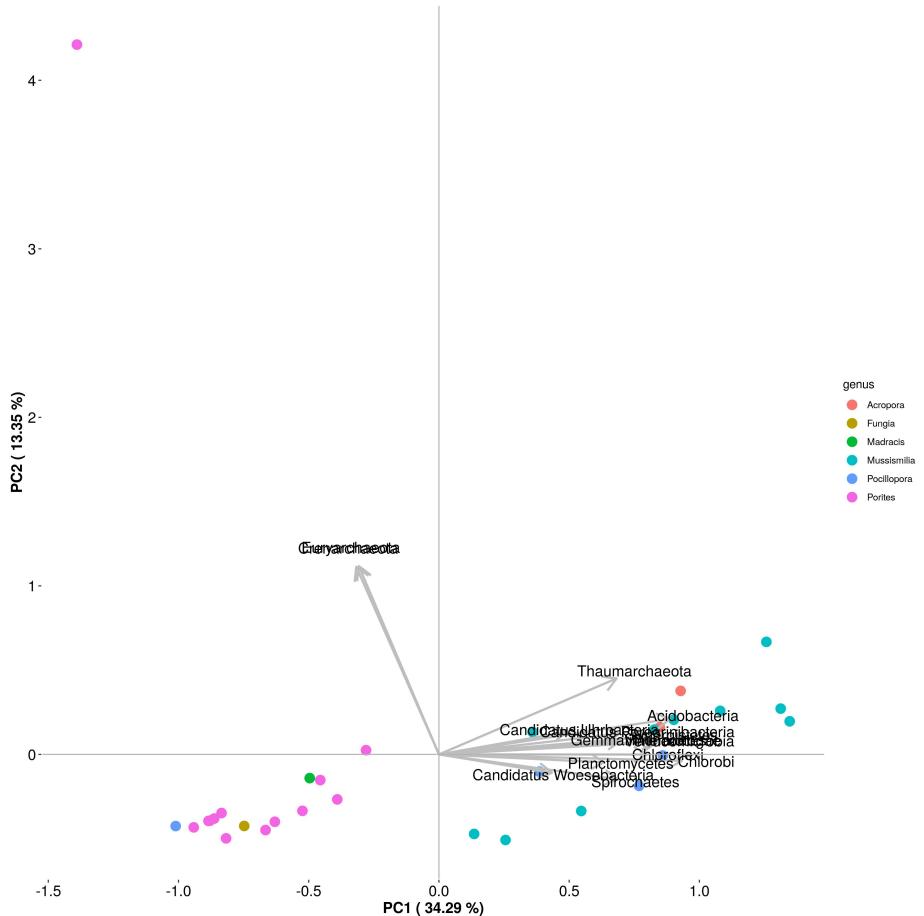


Figura 8.14: PCA GERADO COM OS PRIMEIROS 15 FILOS DO RANDOM FOREST NAO SUPERVISIONADO

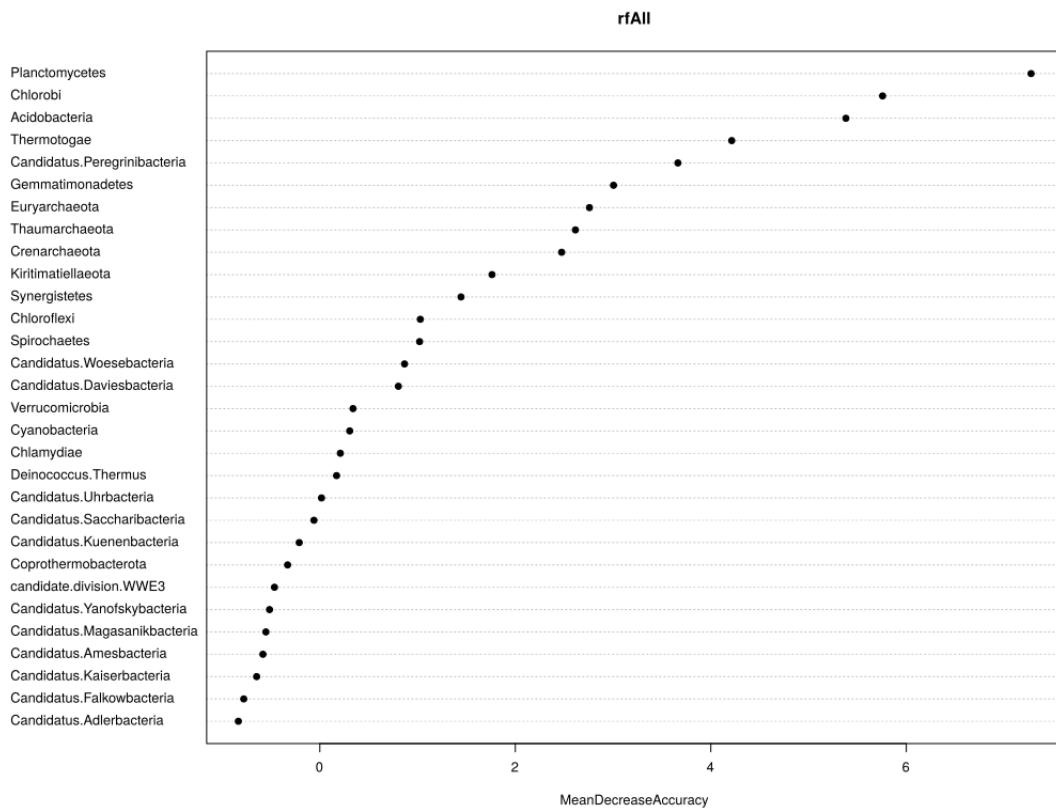


Figura 8.15: RANDOM FOREST SUPERVISIONADO POR GENERO DOS METAGENOMAS DE CORAIS DO MG RAST ANALISADOS COM A BASE DEFINITIVA

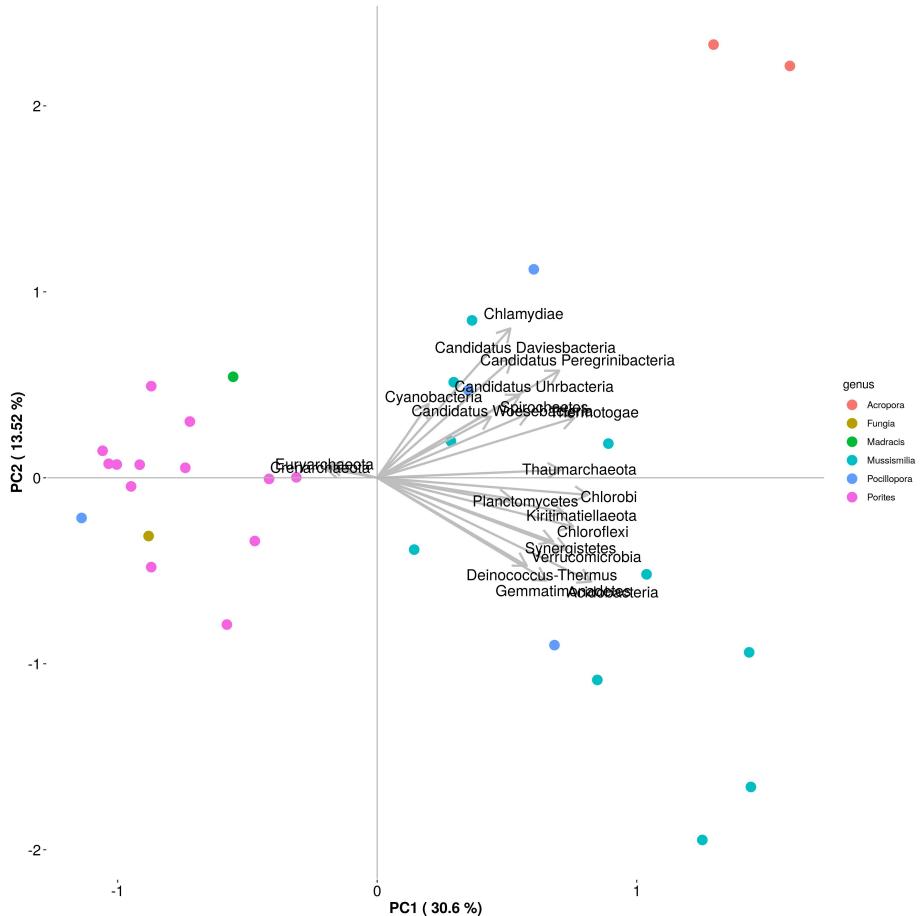


Figura 8.16: PCA GERADO COM OS PRIMEIROS 20 FILOS DO RANDOM FOREST SUPERVISIONADO POR GENERO

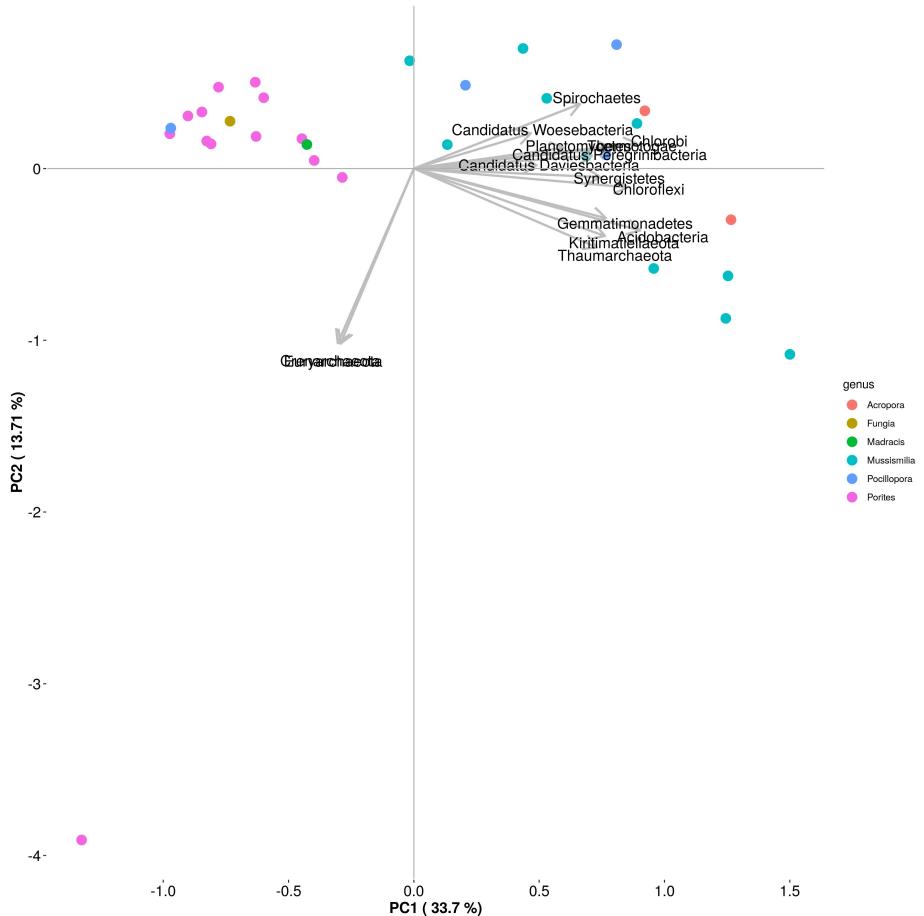


Figura 8.17: PCA GERADO COM OS PRIMEIROS 15 FILOS DO RANDOM FOREST SUPERVISIONADO POR GENERO

8.8 Obtencao de figuras 30/10/2018

O Rilquer sinalizou que a base de dados ainda tinha alguns problemas, entao tive de refazer as analises. Fiz analises para familia e filo. Ordem:

FILO:

- nMDS
- PCA geral
- Random Forest nao supervisionado
- PCA com os 20 primeiros filos indicados pelo RF nao supervisionado
- PCA com os 15 primeiros filos indicados pelo RF nao supervisionado
- PCA com 10 primeiros filos indicados pelo RF nao supervisionado
- Random Forest supervisionado por genero de coral
- PCA com os 20 primeiros filos indicados pelo RF supervisionado por genero

- PCA com os 15 primeiros filos indicados pelo RF supervisionado por genero
- PCA com 10 primeiros filos indicados pelo RF supervisionado por genero
- Random Forest supervisionado por saude de coral
- PCA com os 20 primeiros filos indicados pelo RF supervisionado por saude de coral
- PCA com os 15 primeiros filos indicados pelo RF supervisionado por saude de coral
- PCA com 10 primeiros filos indicados pelo RF supervisionado por saude de coral

FAMILIA:

- nMDS
- PCA geral
- **Random Forest nao supervisionado**
 - PCA com os 20 primeiras familias indicados pelo RF nao supervisionado
 - PCA com os 15 primeiras familias indicados pelo RF nao supervisionado
 - PCA com 10 primeiros familias indicados pelo RF nao supervisionado
- **Random Forest supervisionado por genero de coral**
 - PCA com os 20 primeiros familias indicados pelo RF supervisionado por genero
 - PCA com os 15 primeiros familias indicados pelo RF supervisionado por genero
 - PCA com 10 primeiros familias indicados pelo RF supervisionado por genero
- **Random Forest supervisionado por saude de coral**
 - PCA com os 20 primeiros familias indicados pelo RF supervisionado por saude de coral
 - PCA com os 15 primeiros familias indicados pelo RF supervisionado por saude de coral
 - PCA com 10 primeiros familias indicados pelo RF supervisionado por saude de coral

8.8.1 FILO

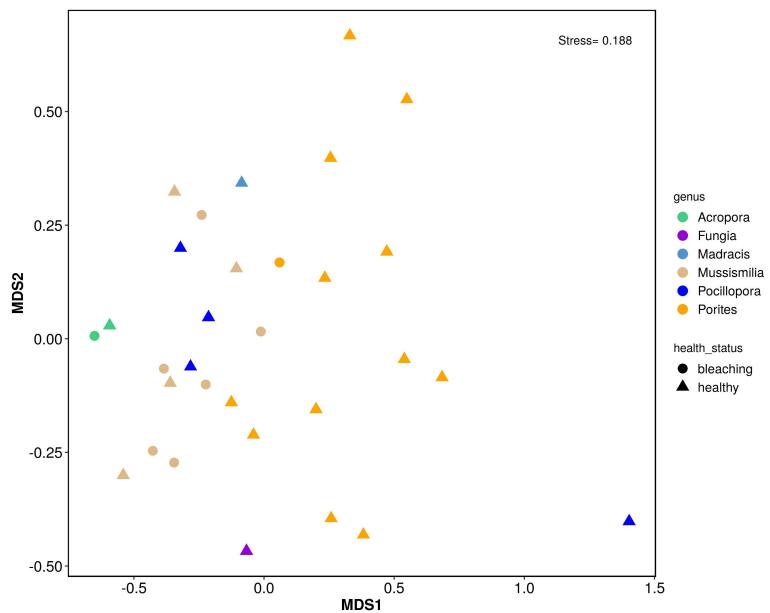


Figura 8.18: nMDS feito com matriz de abundancia de filos

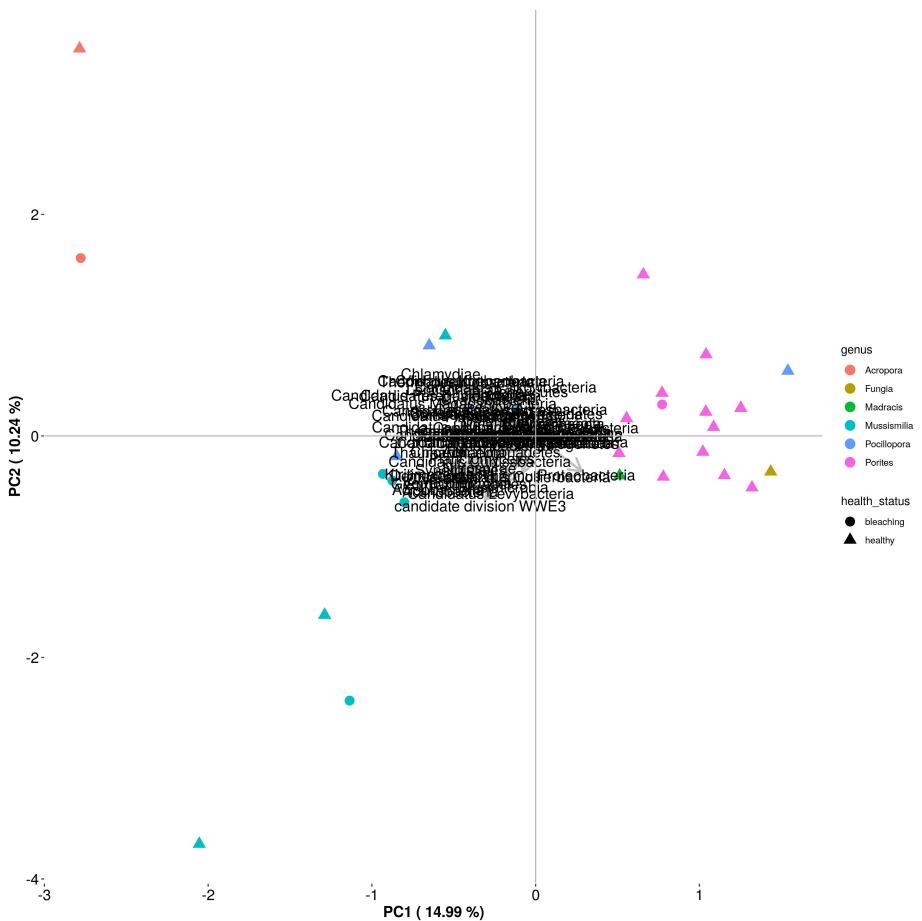


Figura 8.19: PCA geral feito com matriz de abundancia de filos

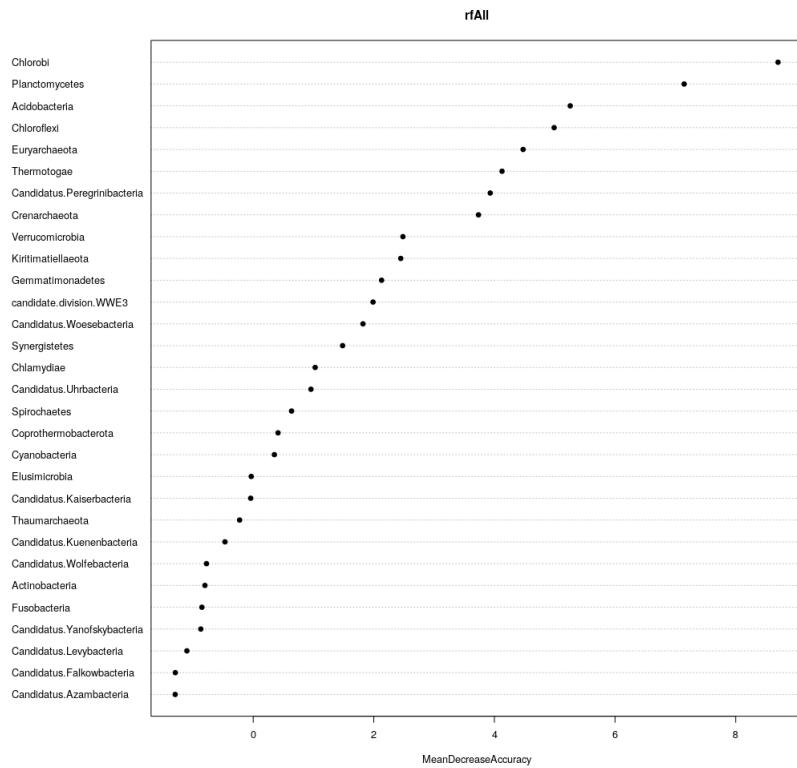


Figura 8.20: RANDOM FOREST NÃO SUPERVISIONADO DAS AMOSTRAS DE CORAIS DO MG RAST FEITO DIA 30 DE OUTUBRO

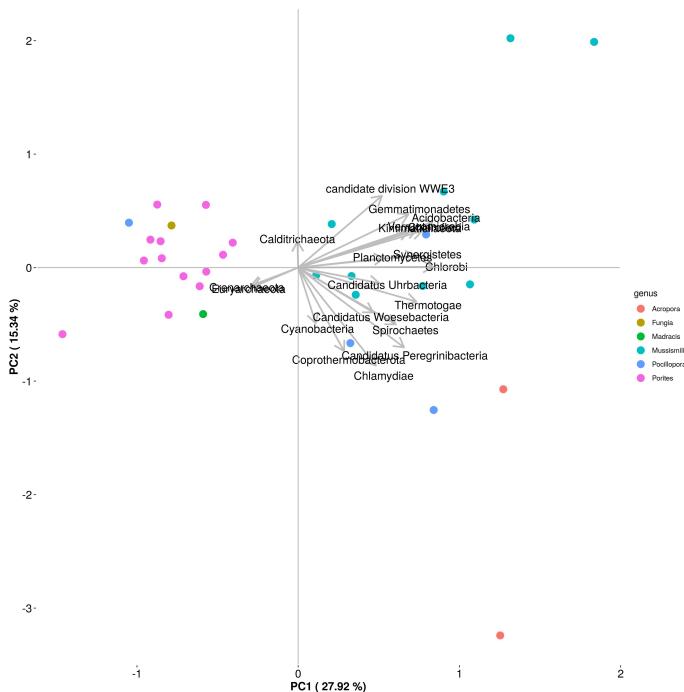


Figura 8.21: PCA feito com os 20 primeiros filos indicados pelo Random Forest não supervisionado 30 DE OUTUBRO

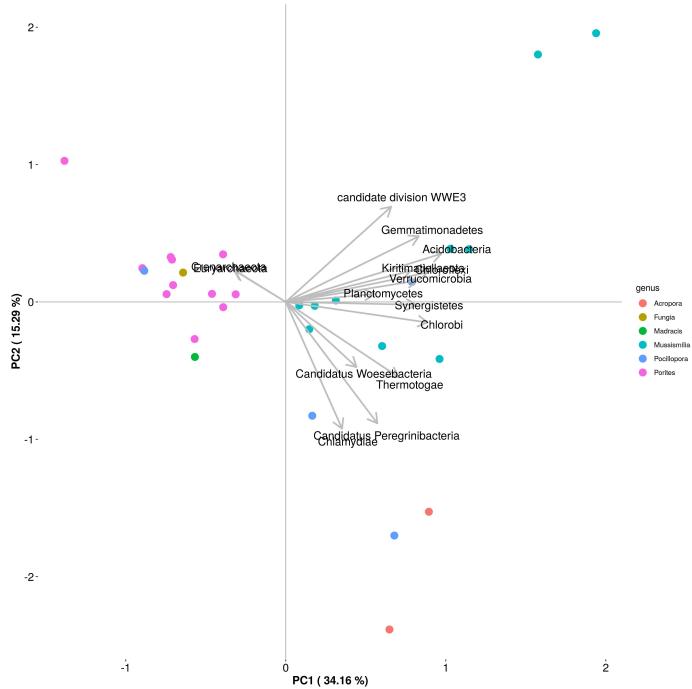


Figura 8.22: PCA feito com os 15 primeiros filos indicados pelo Random Forest não supervisionado 30 DE OUTUBRO

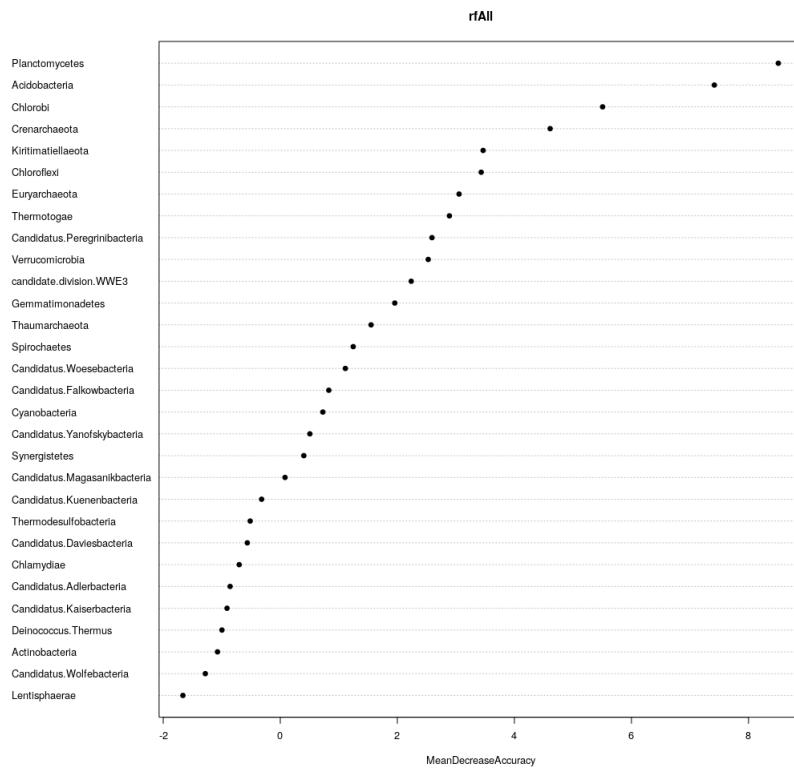


Figura 8.23: Random Forest supervisionado por genero de coral feito 30 de novembro

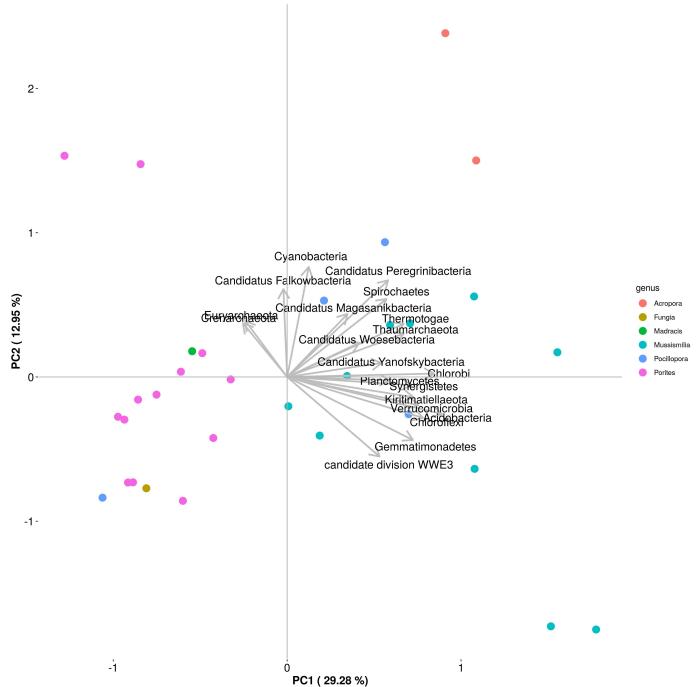


Figura 8.24: PCA com 20 filos indicados pelo Random Forest supervisionado por genero de coral feito 30 de novembro

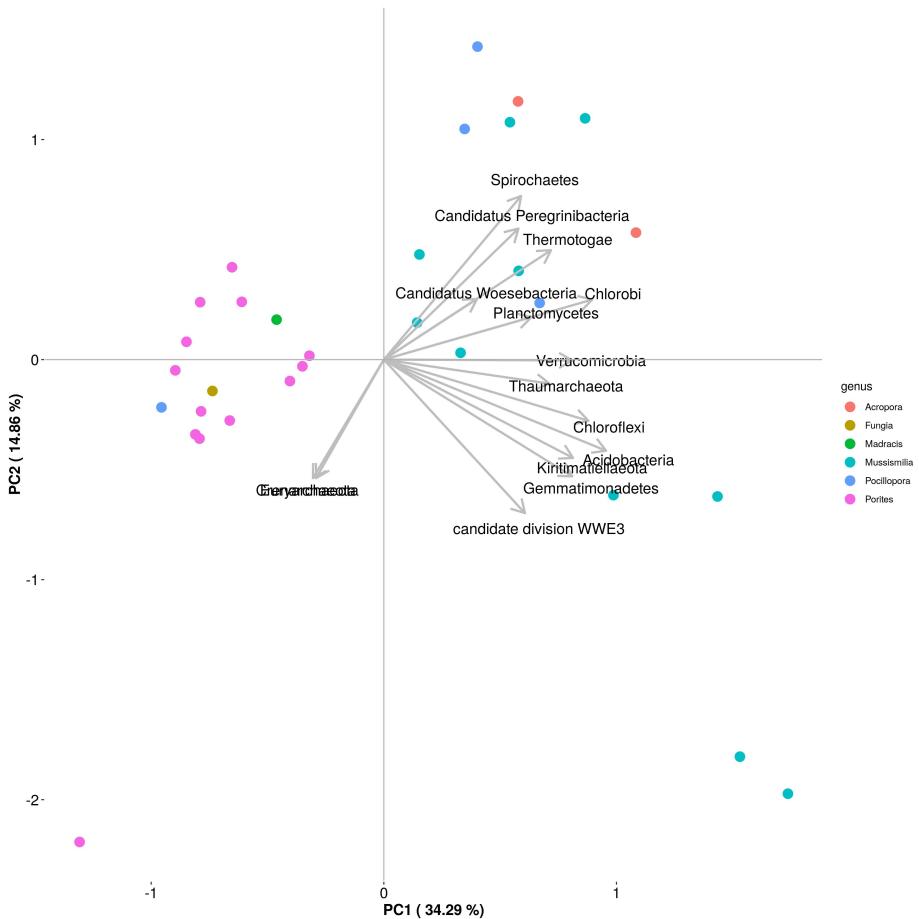


Figura 8.25: PCA com 15 filos indicados pelo Random Forest supervisionado por genero de coral feito 30 de novembro

8.8.2 FAMILIA

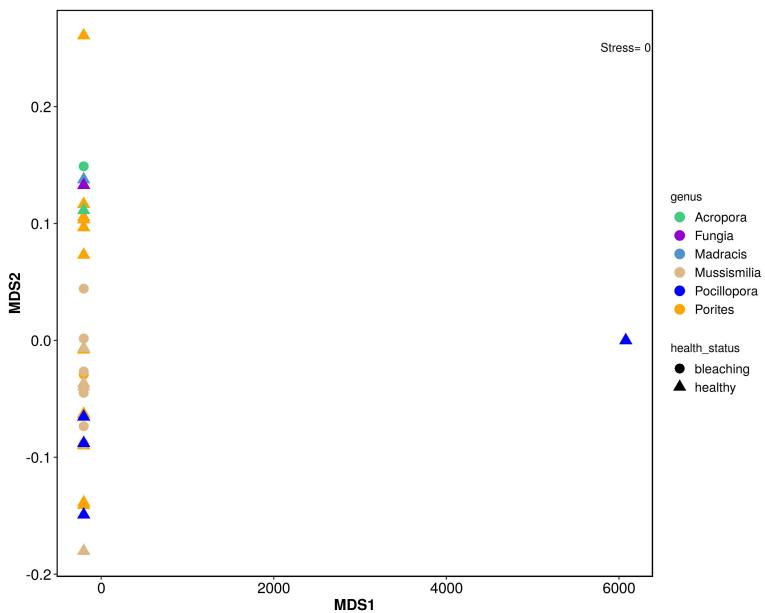


Figura 8.26: nMDS feito com matriz de abundancia de familias

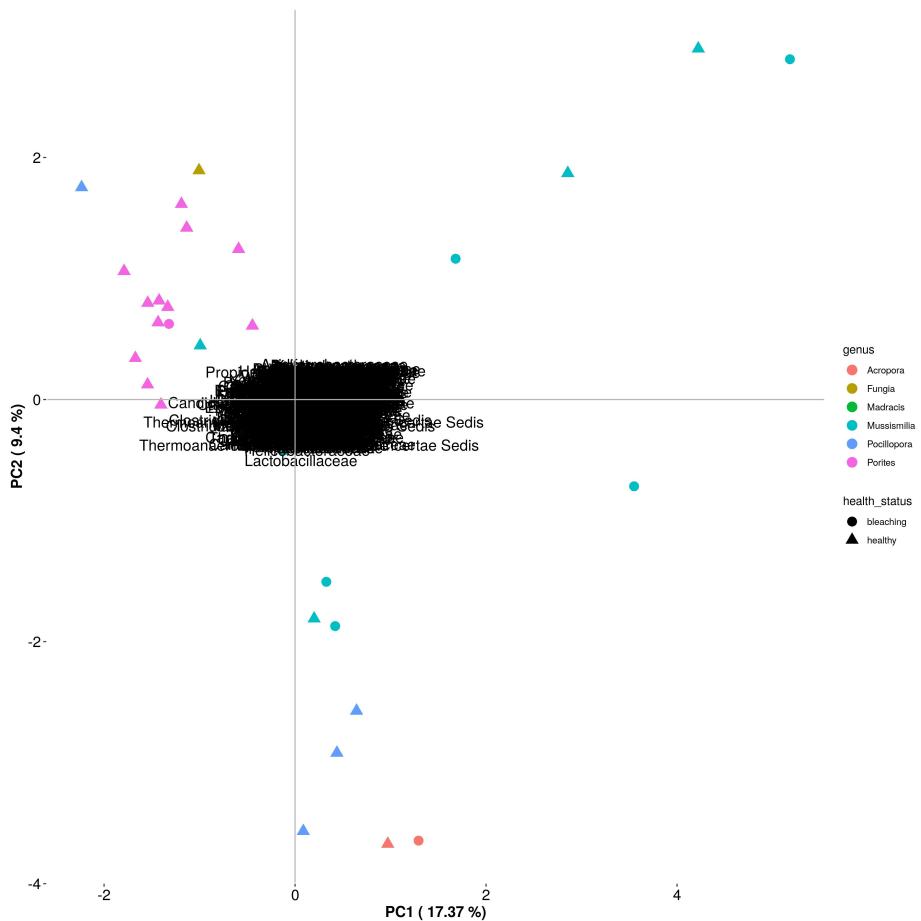


Figura 8.27: PCA geral feito com matriz de abundancia de familias

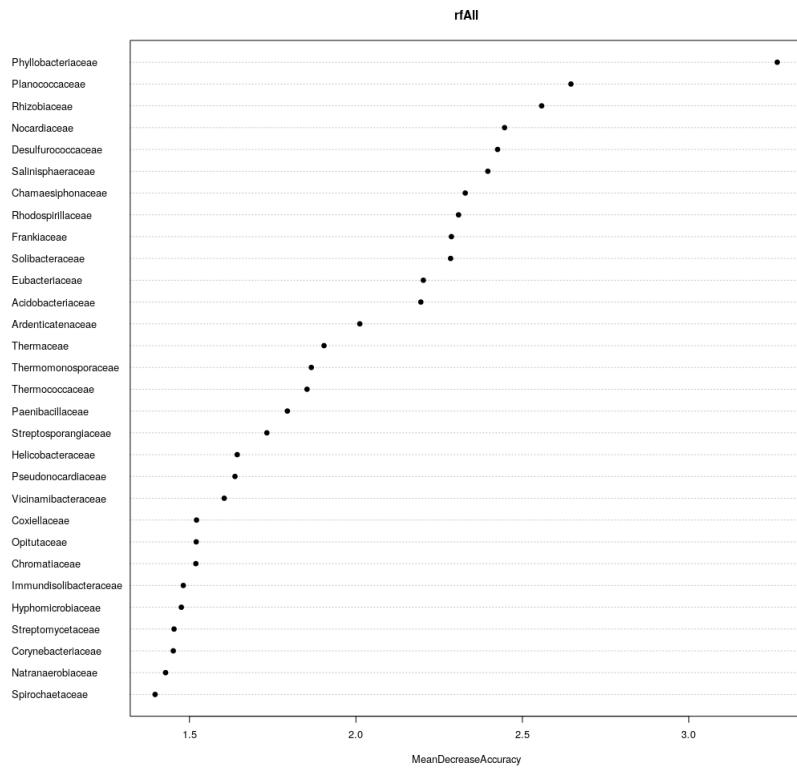


Figura 8.28: RANDOM FOREST NAO SUPERVISIONADO PARA FAMILIAS

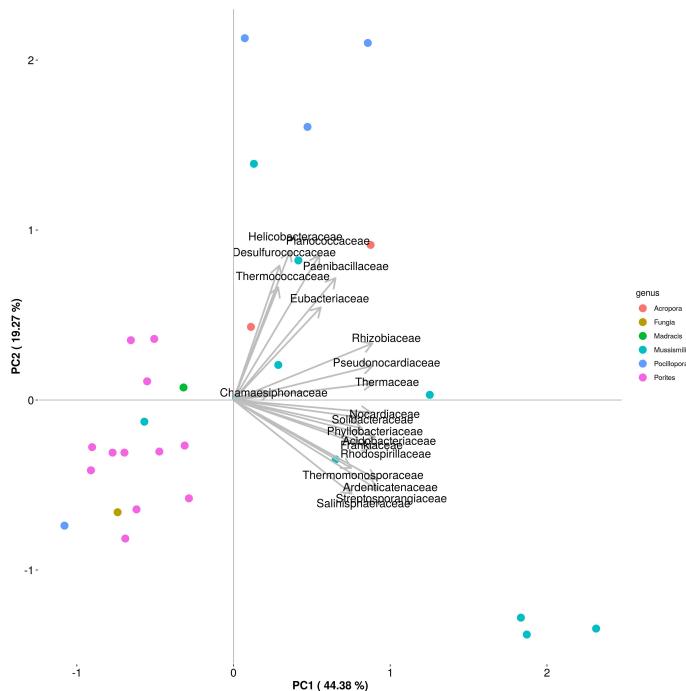


Figura 8.29: PCA FEITO COM OS 20 FAMILIAS INDICADOS PELO RANDOM FOREST NÃO SUPERVISIONADO PARA FAMILIAS

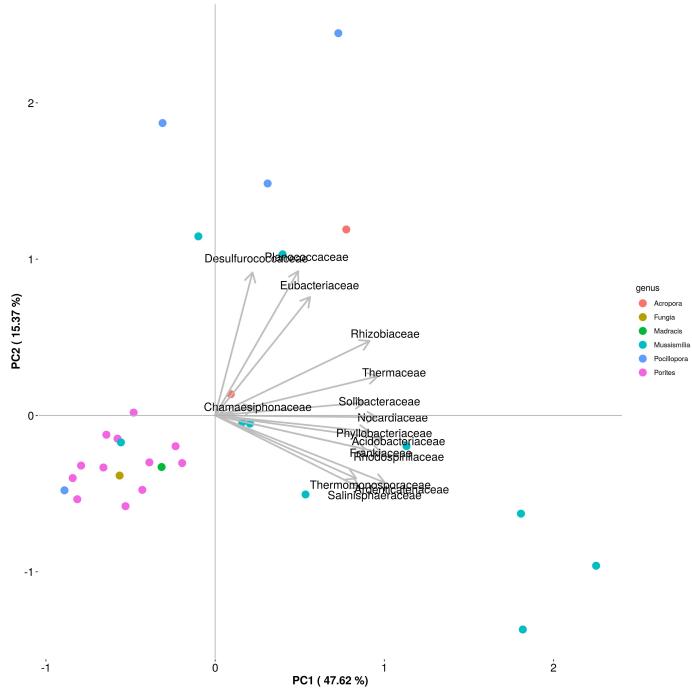


Figura 8.30: PCA FEITO COM OS 15 FAMILIAS INDICADOS PELO RANDOM FOREST NÃO SUPERVISIONADO PARA FAMILIAS

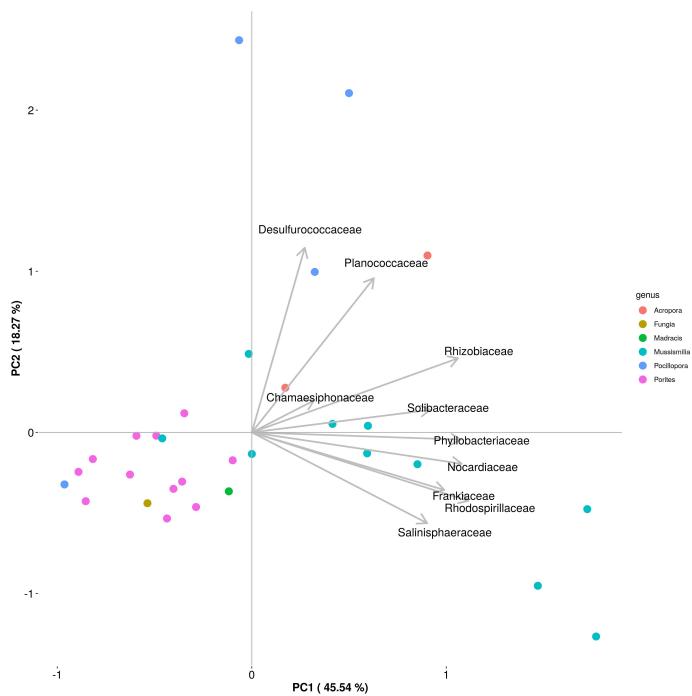


Figura 8.31: PCA FEITO COM OS 10 FAMILIAS INDICADOS PELO RANDOM FOREST NÃO SUPERVISIONADO PARA FAMILIAS

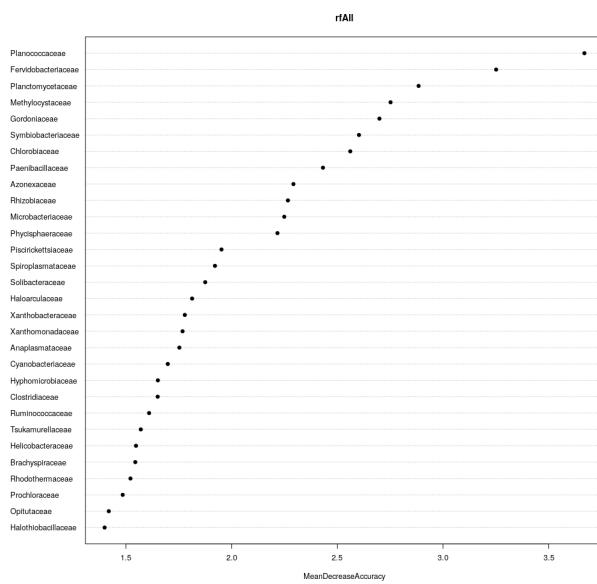


Figura 8.32: RANDOM FOREST SUPERVISIONADO POR GENERO

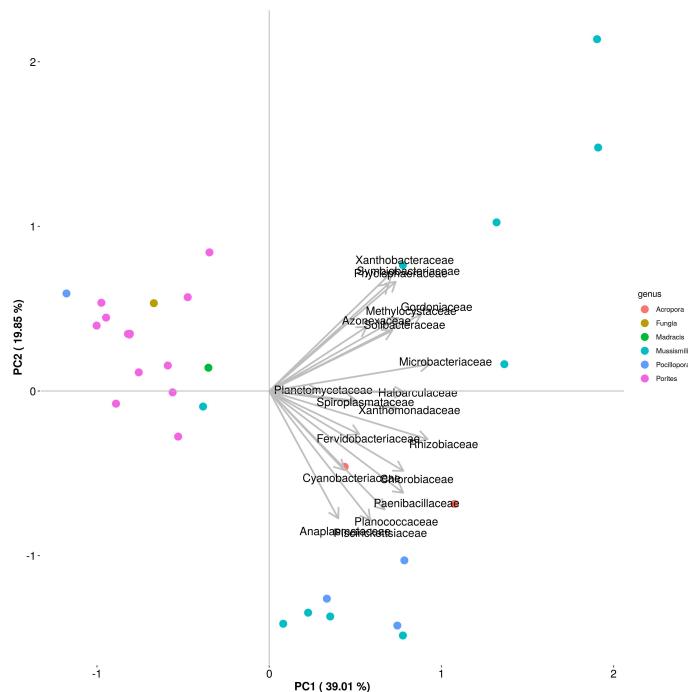


Figura 8.33: PCA FEITO COM OS 20 FAMILIAS INDICADOS PELO RANDOM FOREST SUPERVISIONADO POR GENERO PARA FAMILIAS DE MICRORGANISMOS

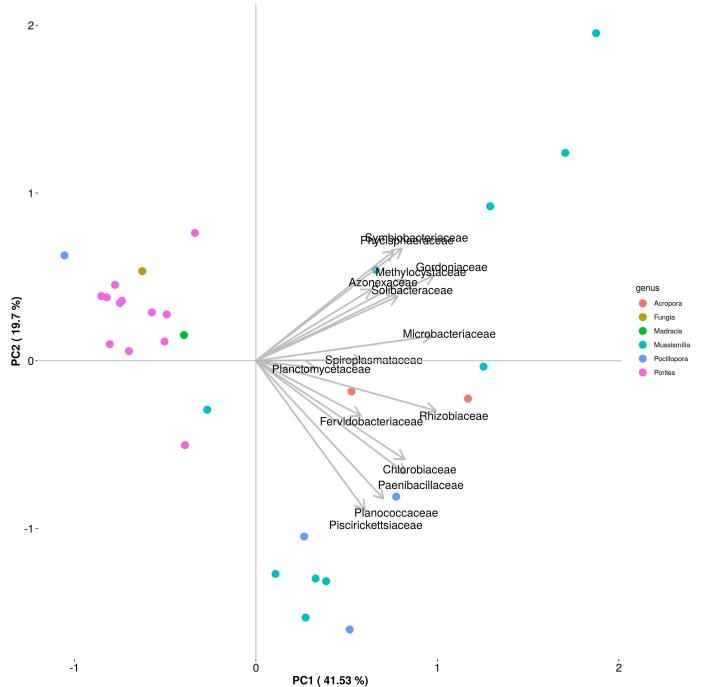


Figura 8.34: PCA FEITO COM OS 15 FAMILIAS INDICADOS PELO RANDOM FOREST SUPERVISIONADO POR GENERO PARA FAMILIAS DE MICRORGANISMOS

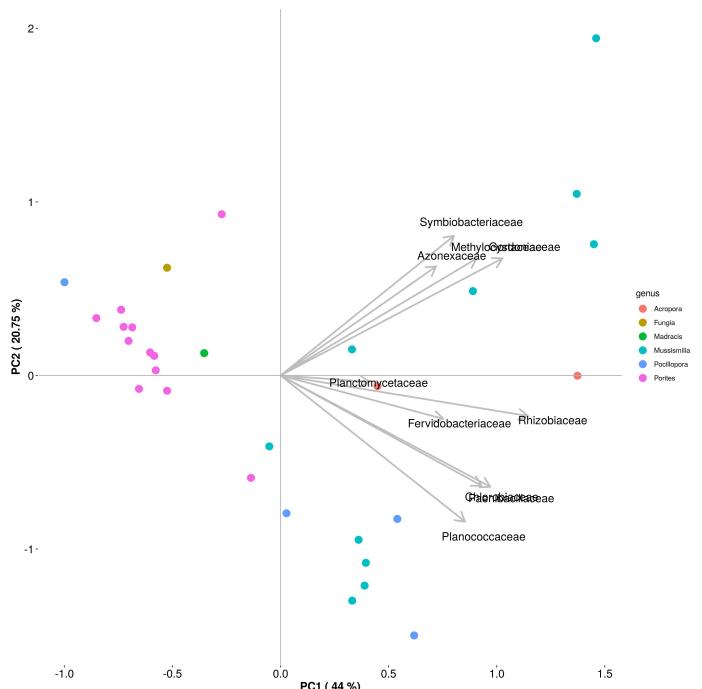


Figura 8.35: PCA FEITO COM OS 10 FAMILIAS INDICADOS PELO RANDOM FOREST SUPERVISIONADO POR GENERO PARA FAMILIAS DE MICRORGANISMOS

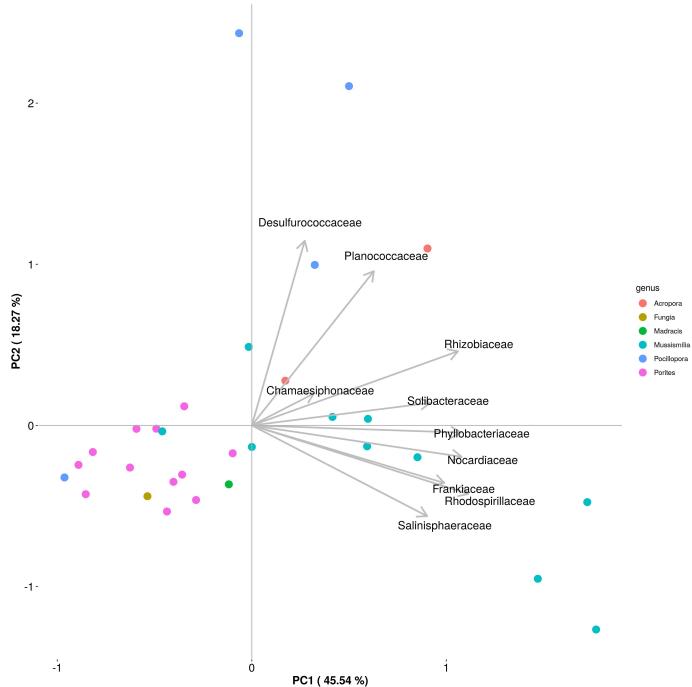


Figura 8.36: RANDOM FOREST SUPERVISIONADO POR SAUDE

8.9 Figuras feitas em dezembro de 2018

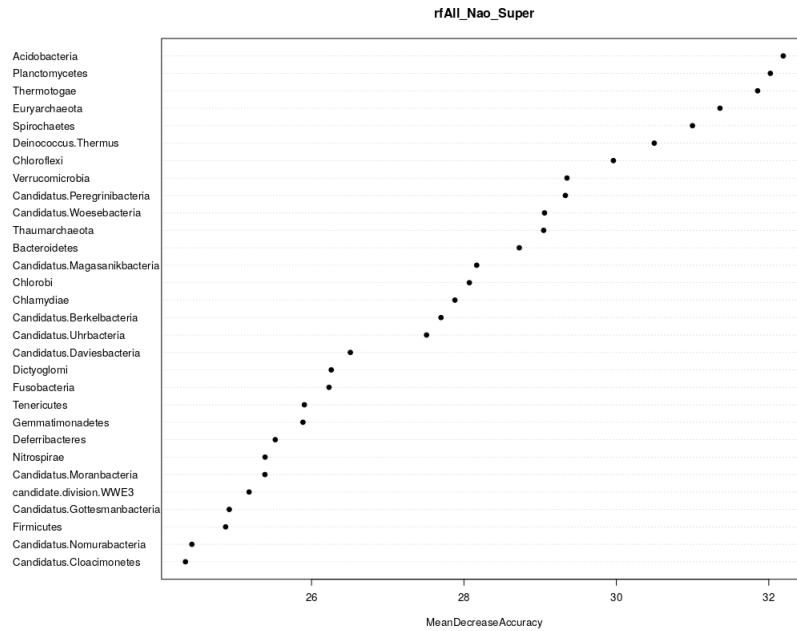


Figura 8.37: RANDOM FOREST NÃO SUPERVISIONADO FEITO DIA 20 DE DEZEMBRO DE 2018

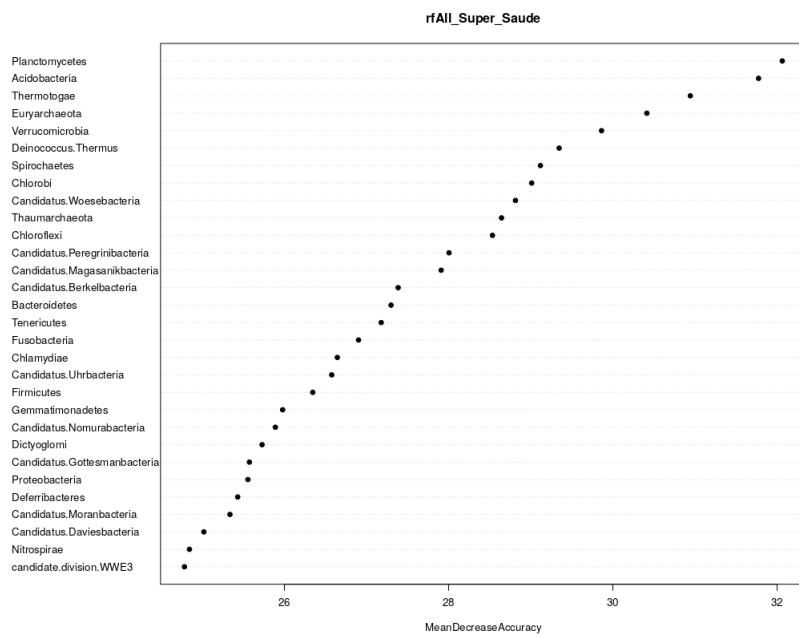


Figura 8.38: RANDOM FOREST SUPERVISIONADO POR ESTADO DE SAUDE FEITO DIA 20 DE DEZEMBRO

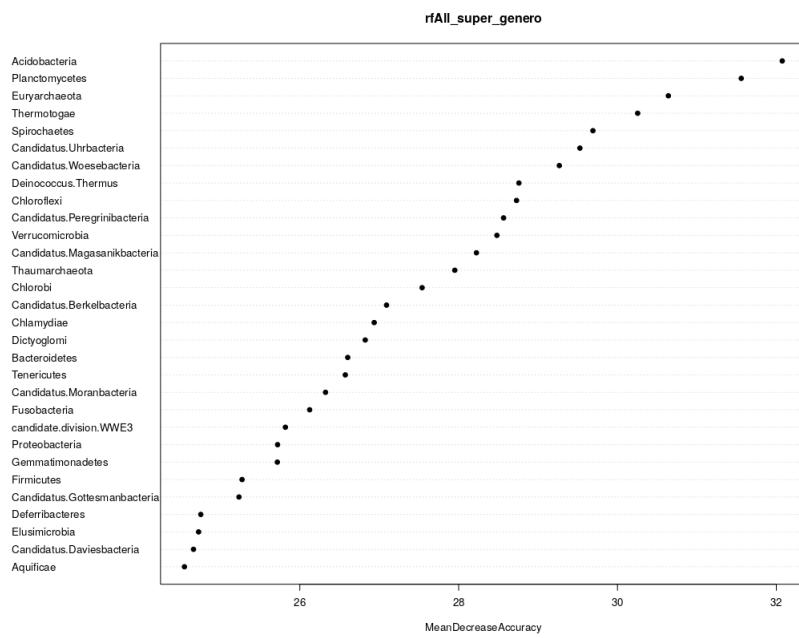


Figura 8.39: RANDOM FOREST SUPERVISIONADO POR GENERO FEITO DIA 20 DE DEZEMBRO

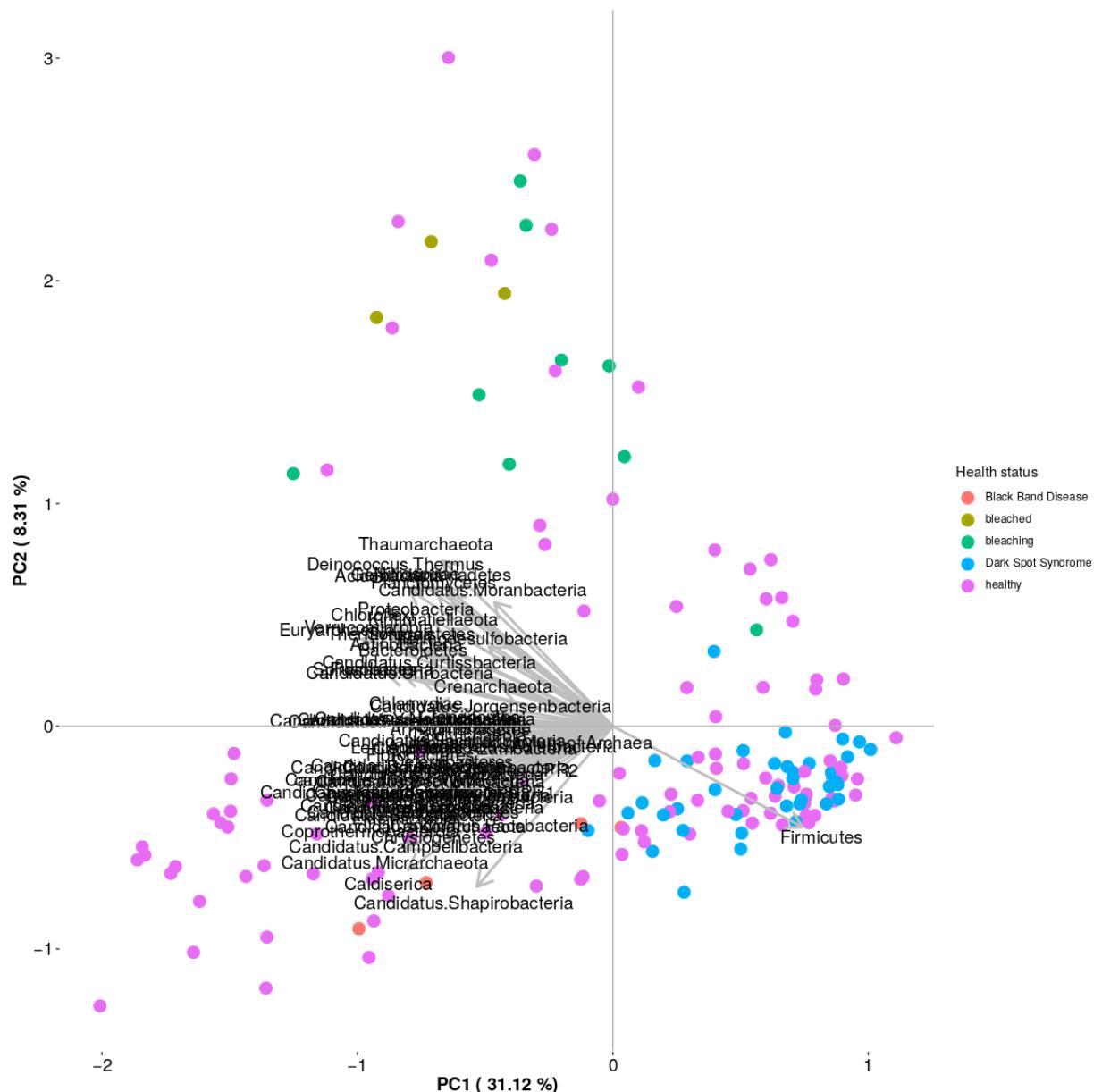


Figura 8.40: PCA FEITO 20 DE DEZEMBRO

Nessa figura, eu noto que diferente dos outros PCAs gerais, existe uma separação muito explícita em vez de existir um grande agrupamento no centro gerado poluicao visual. Sendo que Firmicutes é a variável que direciona a variação das amostras. Não enxergo separação visível entre categorias de estados de saúde.

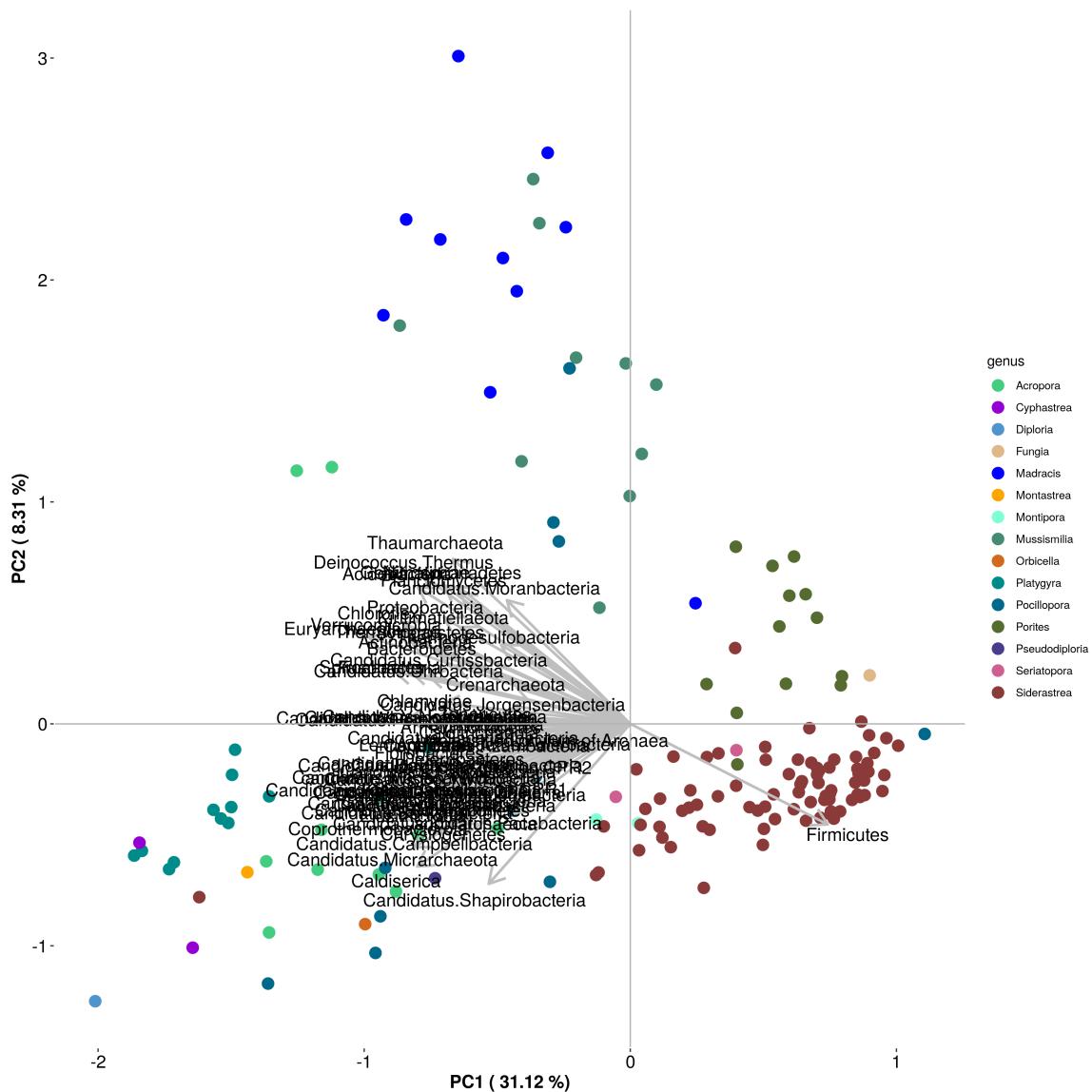


Figura 8.41: PCA GERAL FEITO 20 DE DEZEMBRO VISUALIZANDO GENERO

Nessa figura, enxergo uma separação entre as amostras de diferentes mais clara do que o PCA da página anterior por gênero de corais.

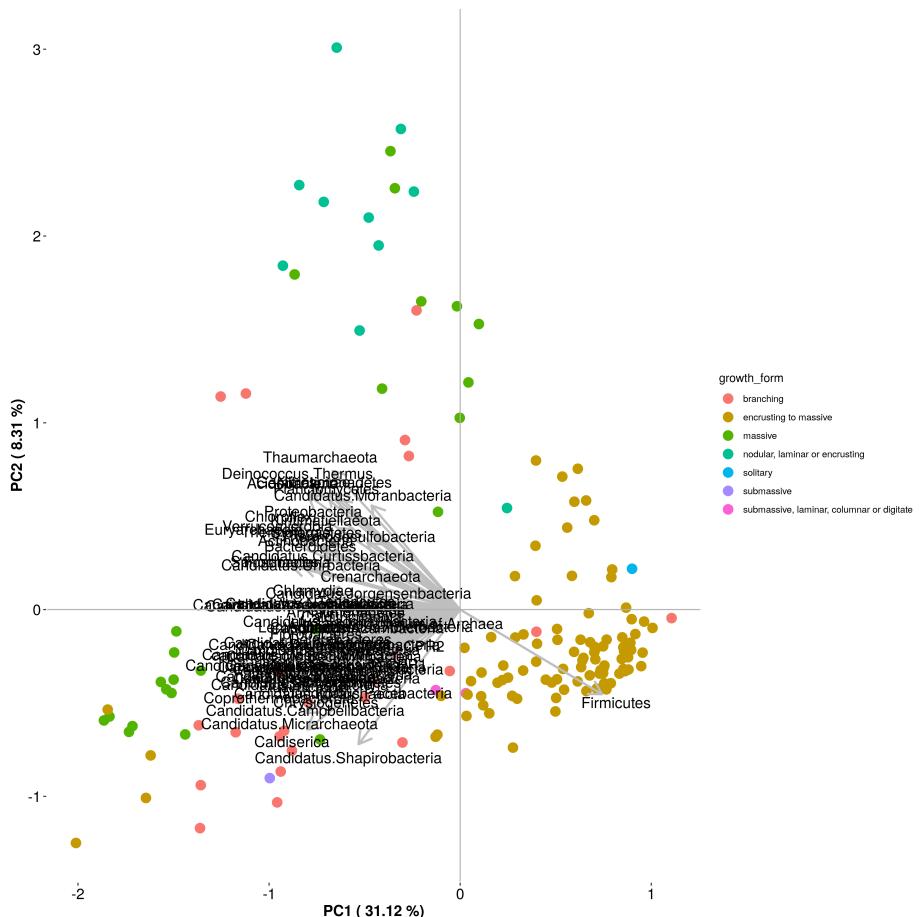


Figura 8.42: PCA GERAL FEITO 20 DE DEZEMBRO VISUALIZANDO CRESCIMENTO

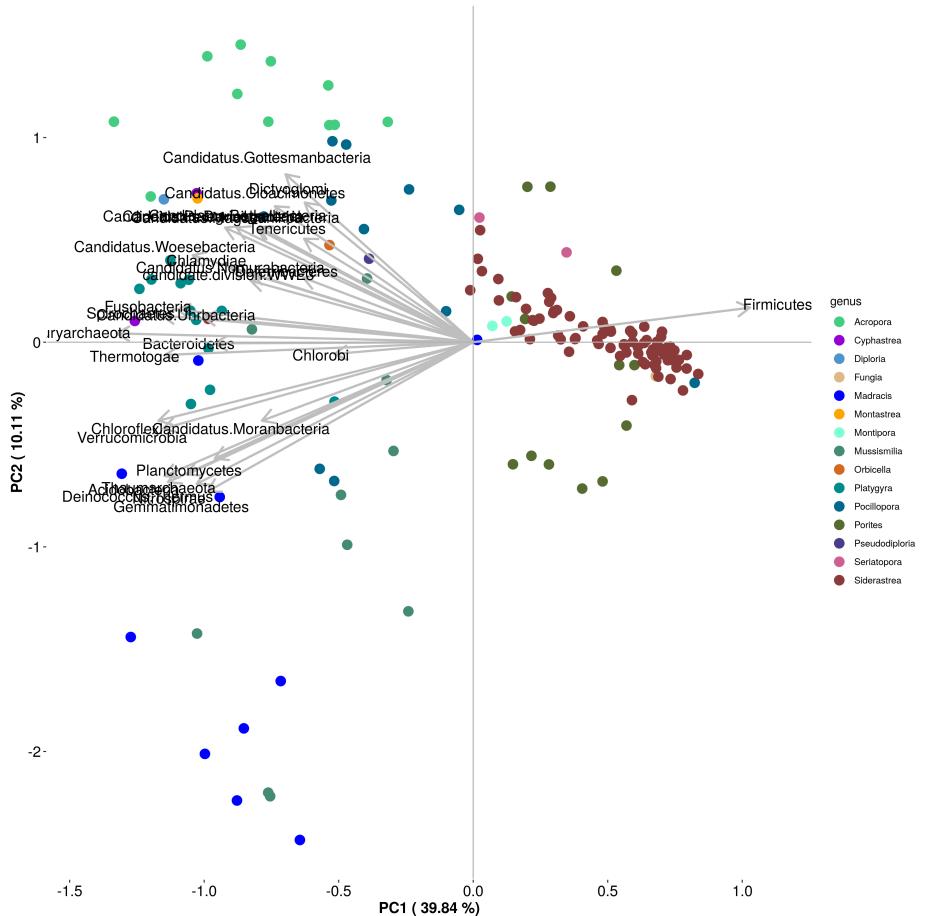


Figura 8.43: PCA COM OS 30 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO CRESCIMENTO

Nessa figura também enxergo uma separação melhor do que por estado de saúde. Entretanto, devo ressaltar que essa melhor separação que enxergo é pela similaridade das amostras do trabalho da Thurber, em que ela trabalha com Dark spot syndrome.

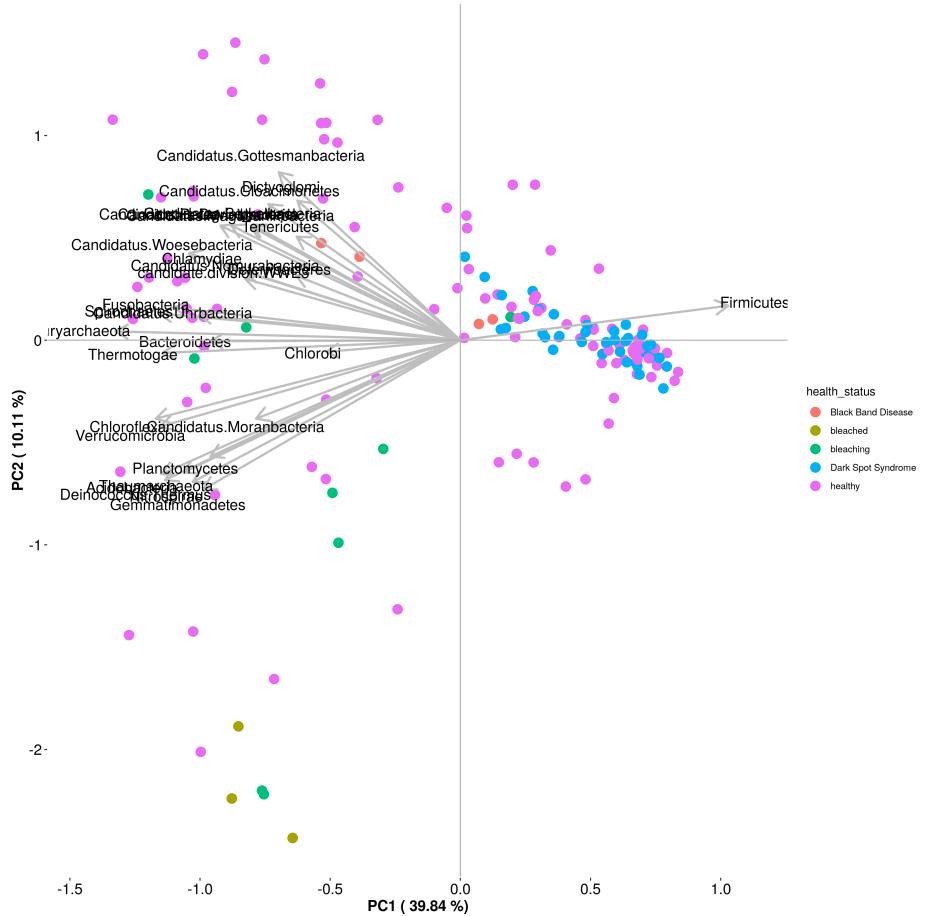


Figura 8.44: PCA COM OS 30 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO SAÚDE

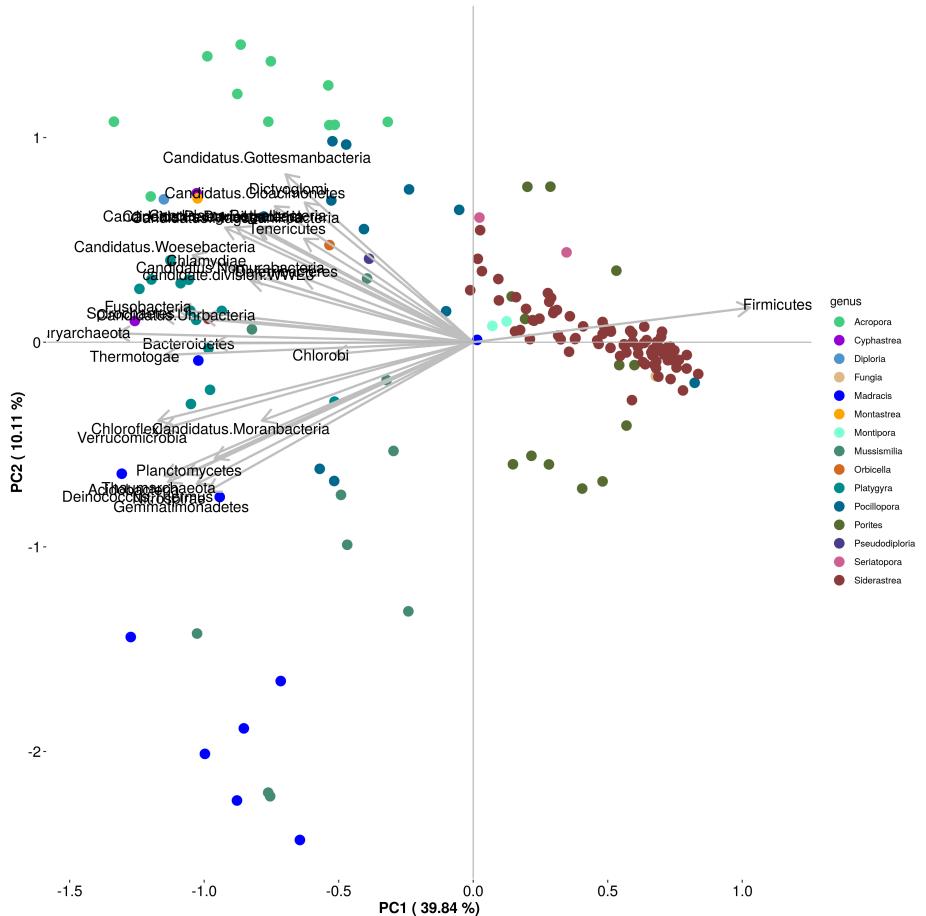


Figura 8.45: PCA COM OS 30 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO GENERO

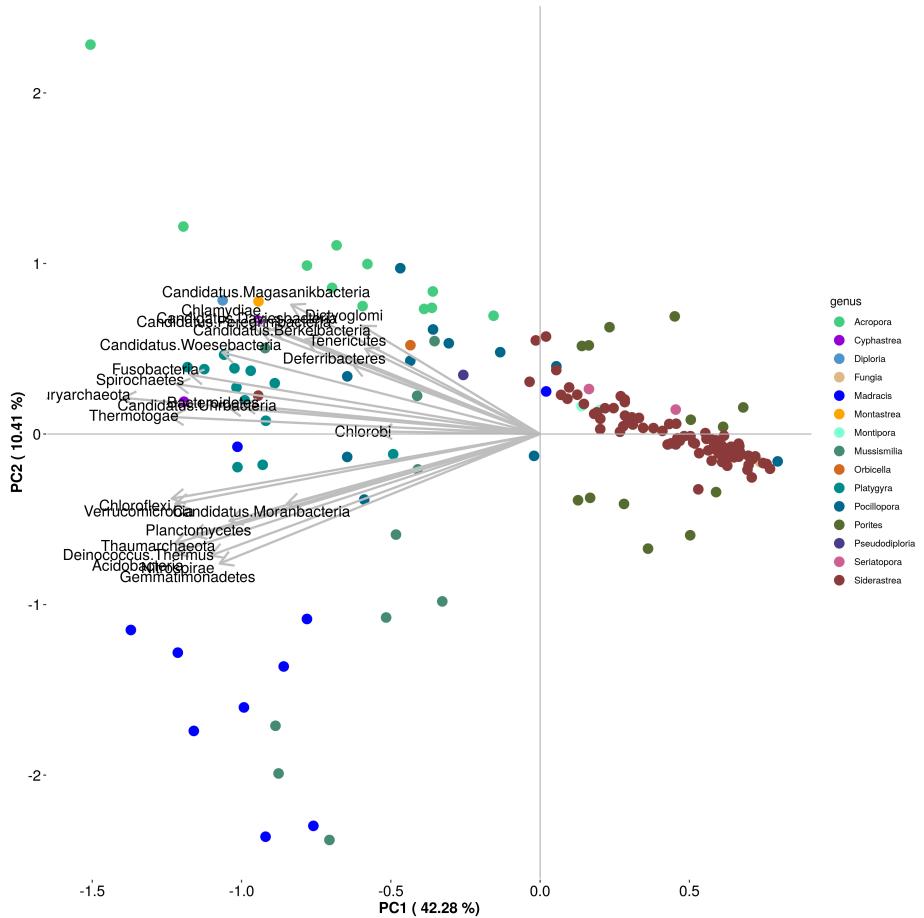


Figura 8.46: PCA COM OS 25 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO GENERO

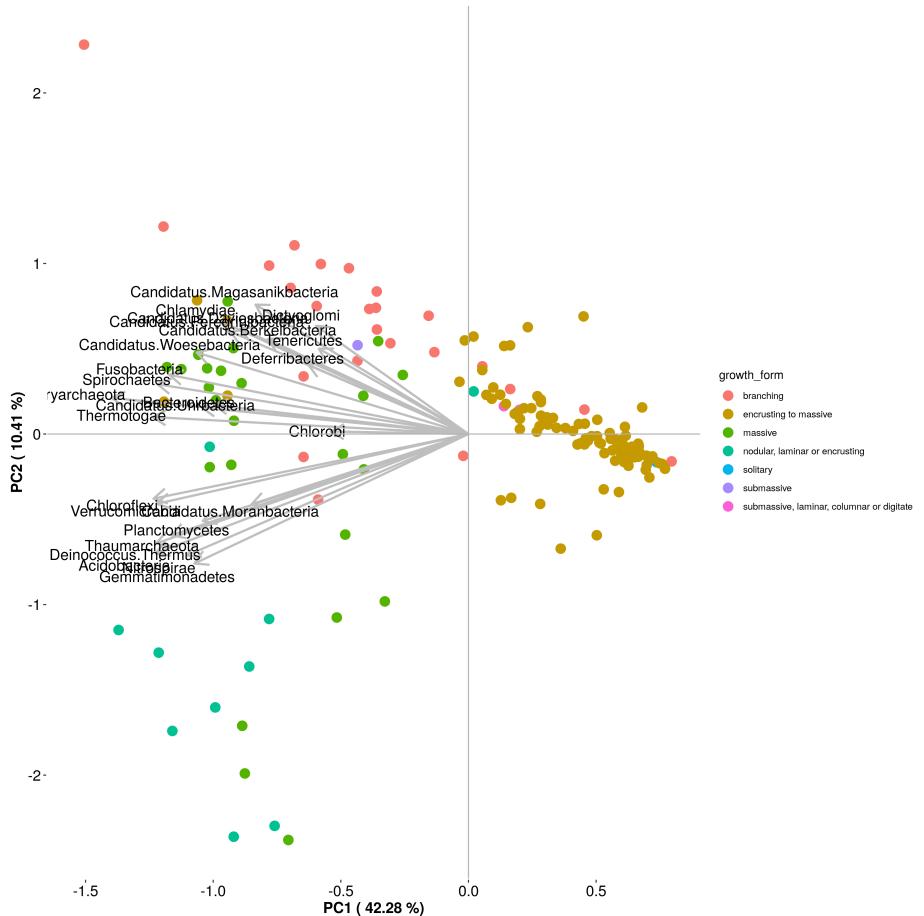


Figura 8.47: PCA COM OS 25 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO CRESCIMENTO

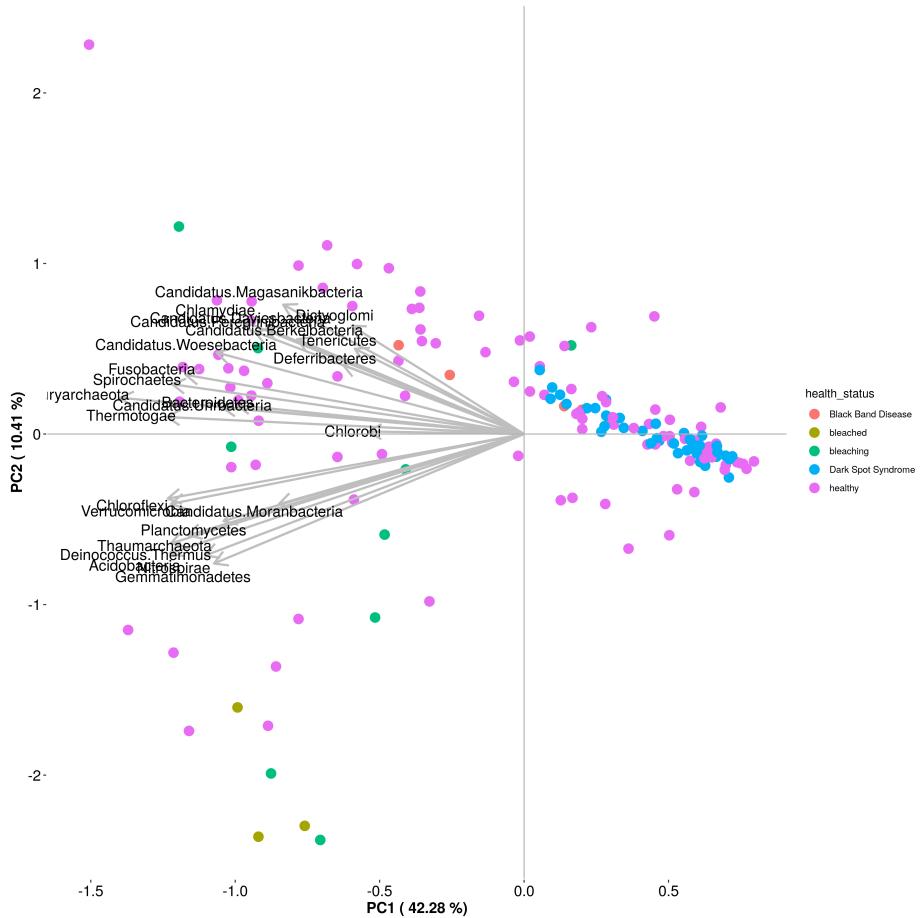


Figura 8.48: PCA COM OS 25 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO SAUDE

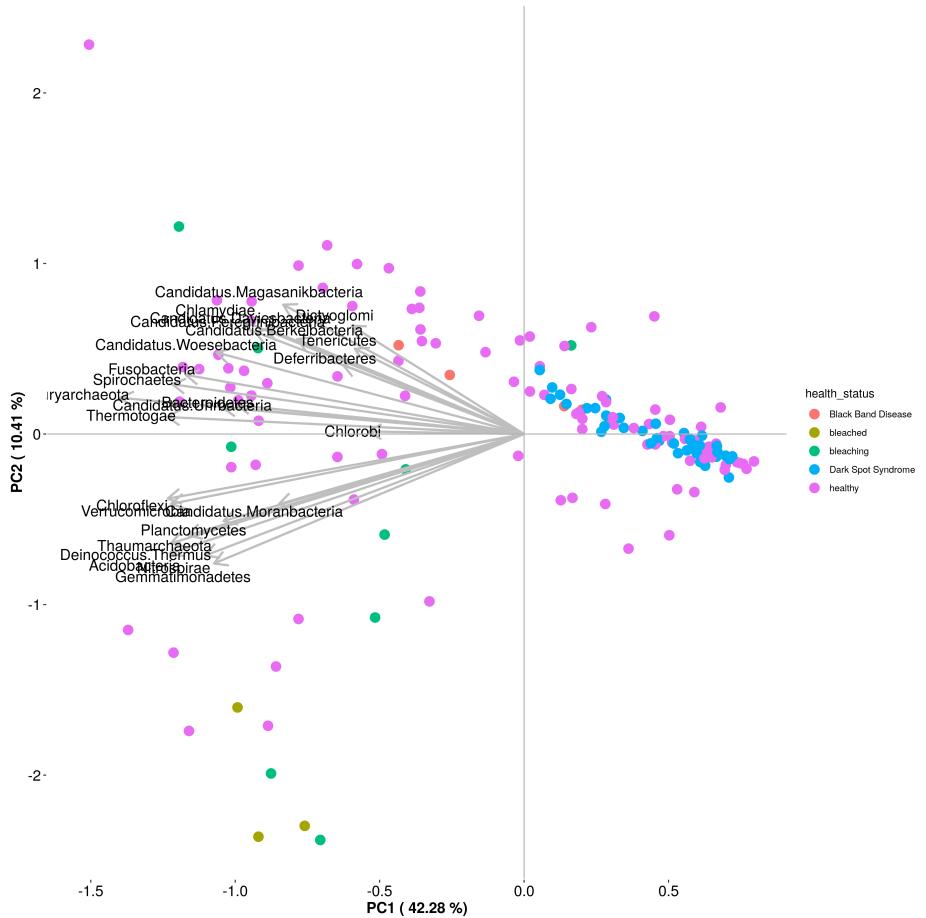


Figura 8.49: PCA COM OS 25 FILOS INDICADOS COMO MAIS RELEVANTES PARA CLASSIFICAR AS AMOSTRAS PELO RANDOM FOREST NÃO SUPERVISIONADO FEITO 20 DE DEZEMBRO VISUALIZANDO SAUDE

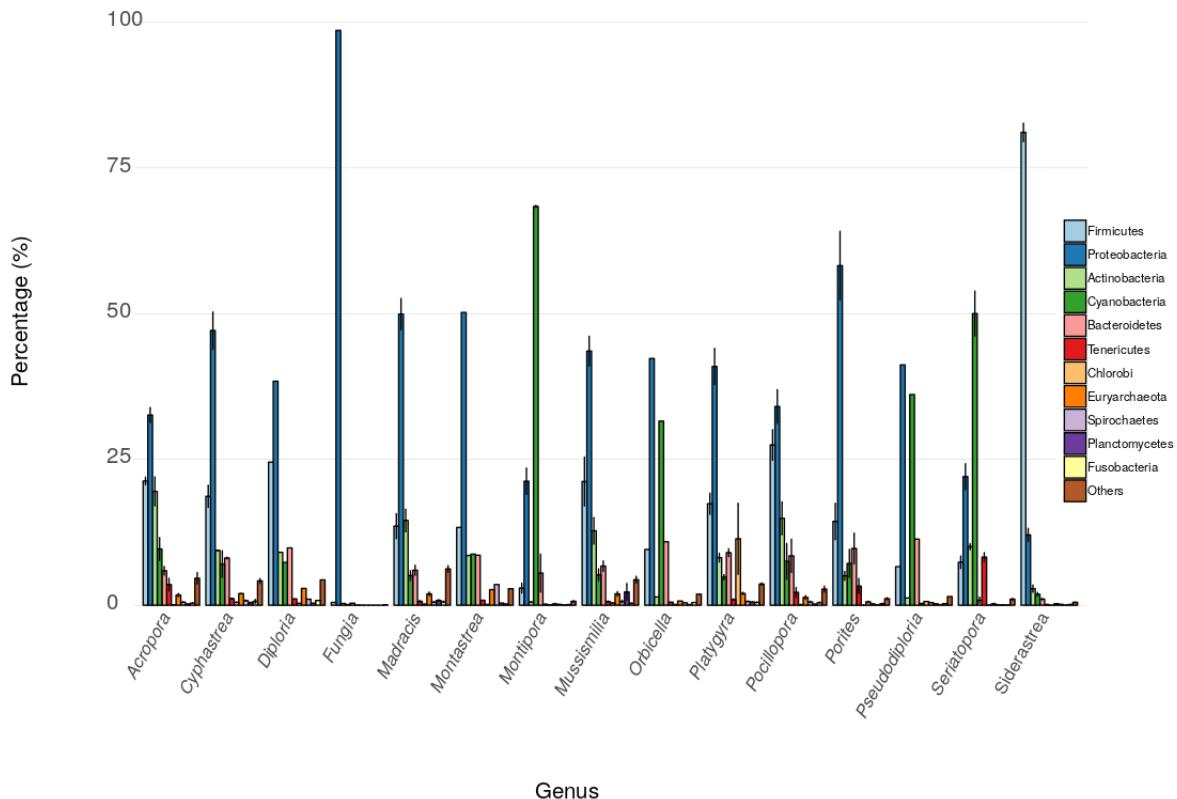
8.10 Fevereiro de 2019

As figuras a seguir ficarão obsoletas com a anotação que botei para rodar dia 05/02.

8.10.1 Graficos de abundancia

Graficos com categorias isoladas

Genero



Genus

Figura 8.50: Grafico de abundancia dos 12 filos mais abundantes por genero de corais

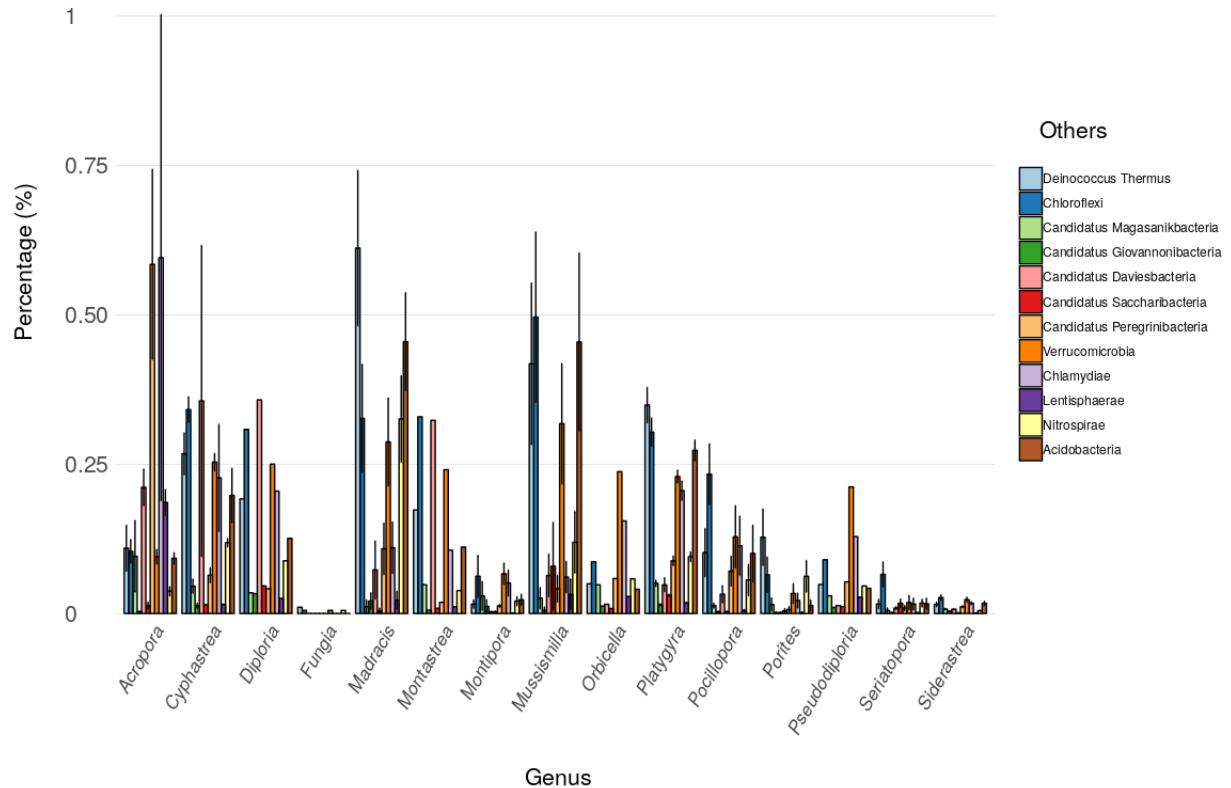


Figura 8.51: Grafico de abundancia dos 12 filos mais abundantes do agrupamento Others por genero de coral

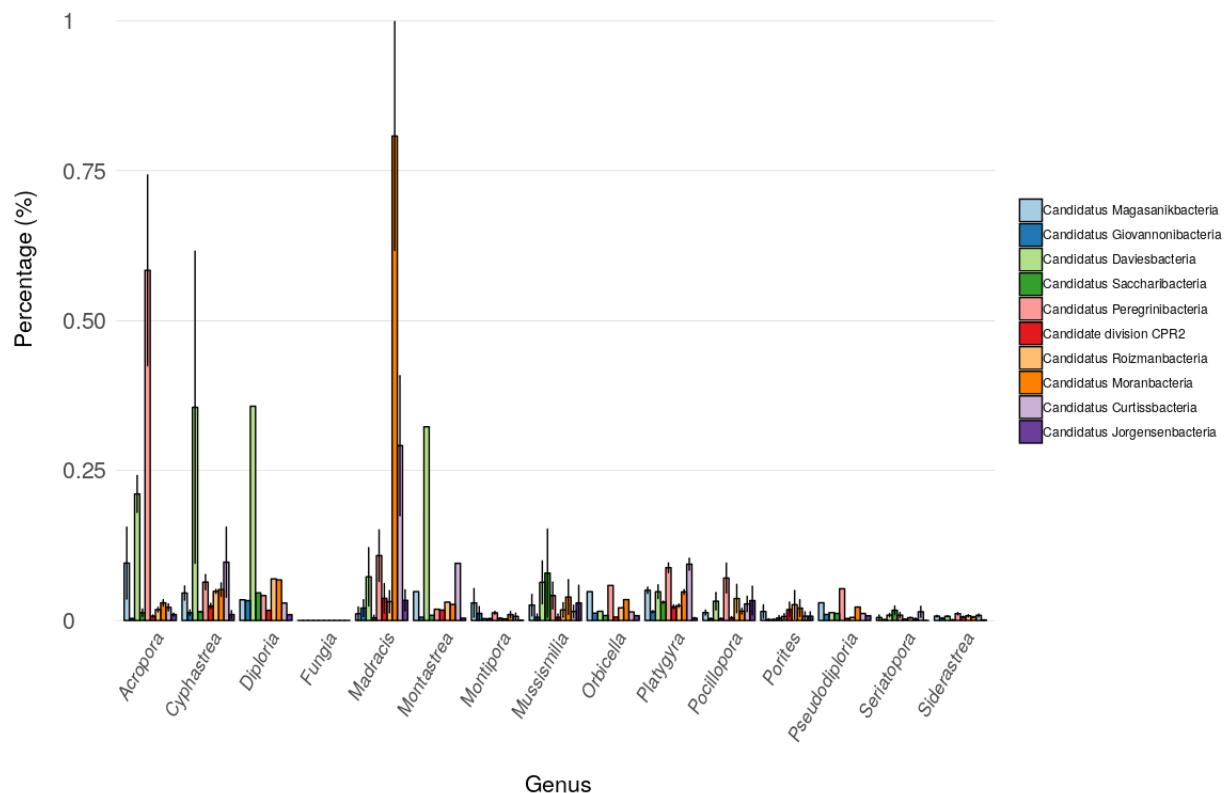


Figura 8.52: Grafico de abundancia dos 10 filos candidatos mais abundantes do agrupamento Others por genero de coral

Graficos com categorias mistas

Genero e Health status

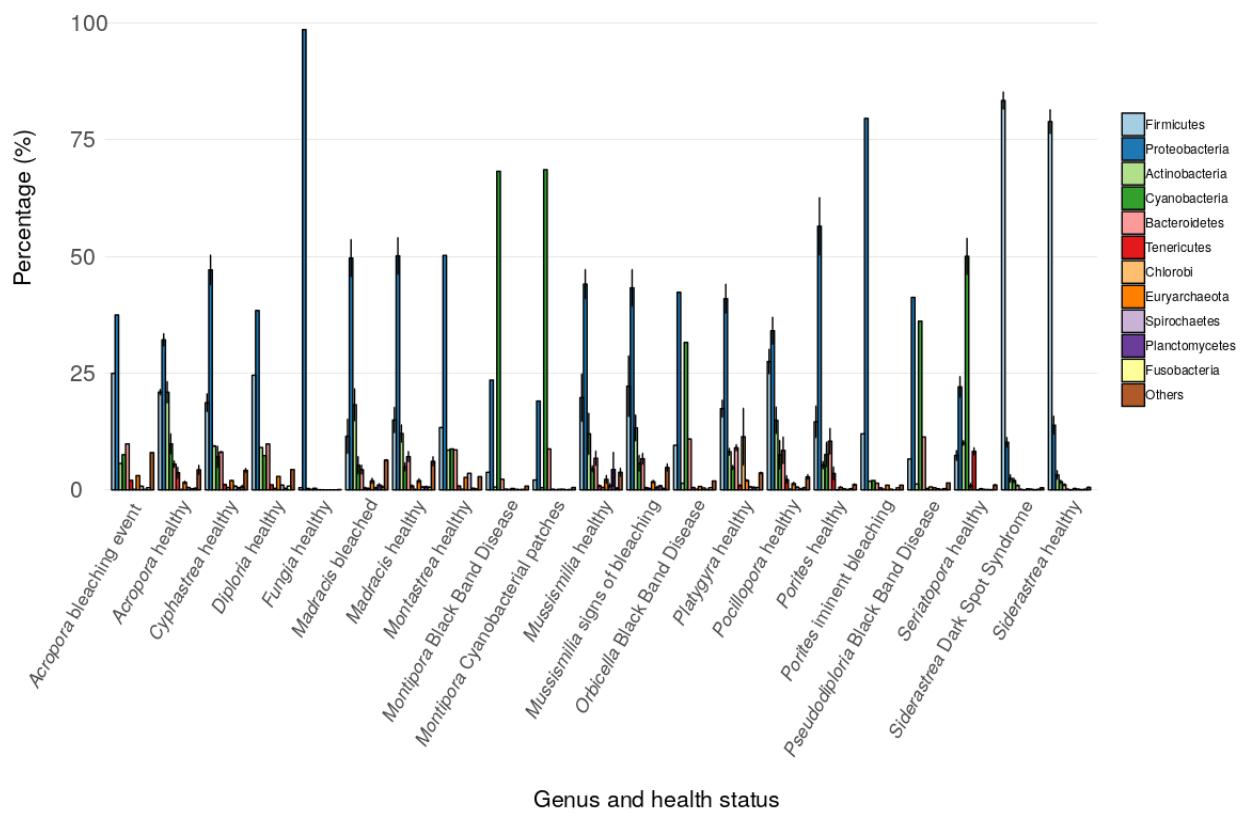


Figura 8.53: Grafico de abundancia dos 12 filos mais abundantes por genero e estado de saude de corais

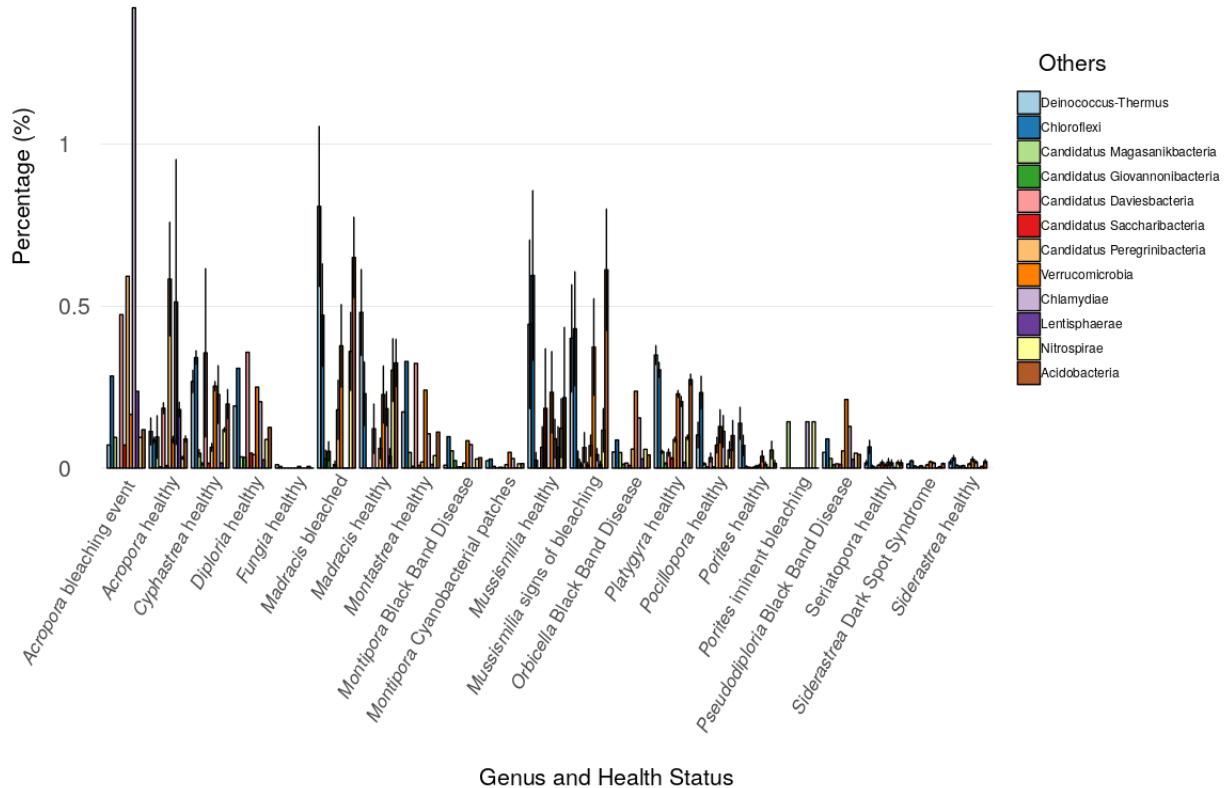


Figura 8.54: Grafico de abundancia dos 12 filos mais abundantes do agrupamento Others por genero e estado de saude do coral

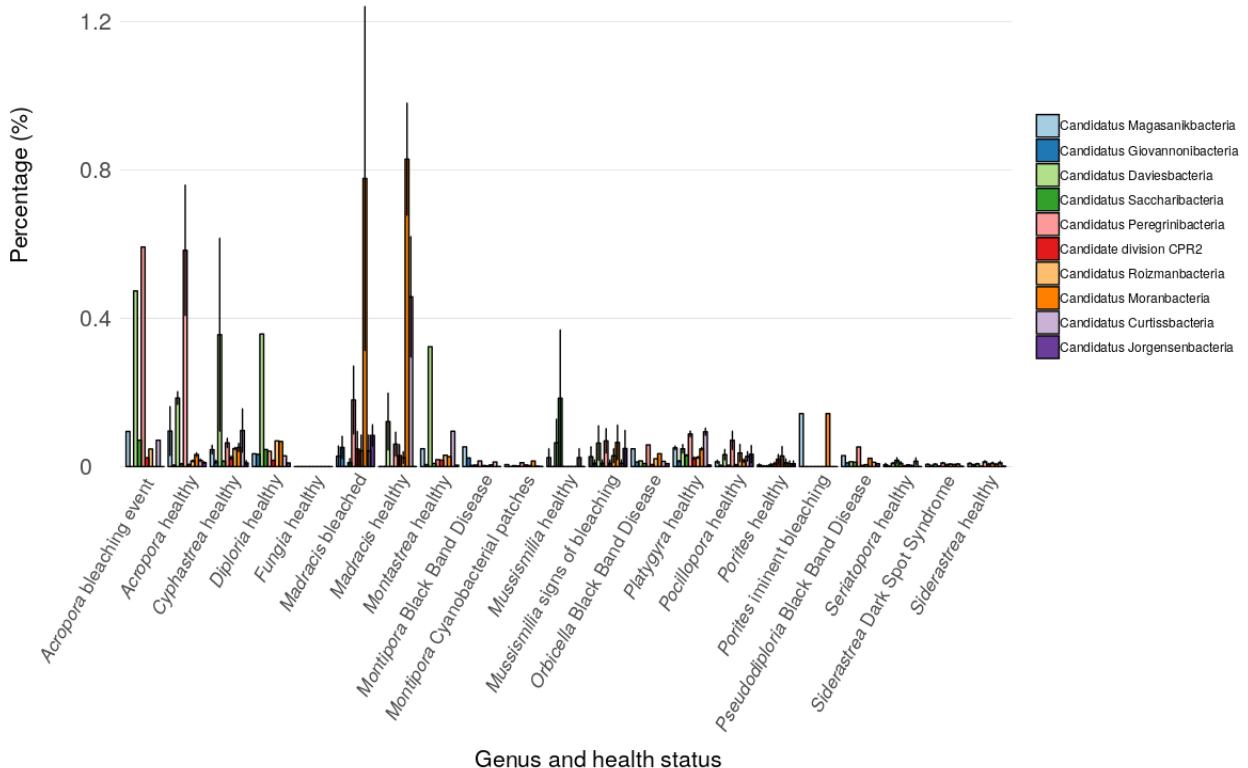


Figura 8.55: Grafico de abundancia dos 10 filos candidatos mais abundantes do agrupamento Others por genero e estado de saude de coral

Eu estou fazendo figuras demais e honestamente não sei mais o que devo avaliar. Isso ficou evidente na reunião do dia 15/02. Amaro e Miguel sugeriram uma leitura teórica maior e uma reflexão sobre o objetivo do meu trabalho, para que eu não tenha que fazer tantas figuras.

8.10.2 Problemas no Atlântico

A algumas semanas atrás, quando o Rilquer falou que a base de dados do Kraken já estava pronta, eu resolvi colocar as anotações do Kraken para rodar. Depois de dois dias na fila, notei que, embora o status do job na fila seja 'R', de "running", o tempo de uso não saia do 00:00:00. E não havia output algum na pasta. Relatei esse problema ao Rilquer e ao Pedro. Duas possibilidades levantadas: problemas no job ou problemas no Atlântico, sendo que um não excluía o outro. O Rilquer avaliou a possibilidade de ser um problema no Atlântico e relatou que consertou o script do TAXONPROFILING, pois esse script não estava reconhecendo arquivos limpos de metagenomas. Entretanto, depois de consertar o script, o problema permaneceu o mesmo. Relatei isso ao Rilquer e ele decidiu fazer um teste com metagenomas de solo e ele teve o mesmo problema. Decidi rodar o kraken isoladamente no 'head node', utilizando bash e nohup. Para as amostras do MG-RAST, deu certo e os reports atualizados do mg-rast já estão prontos. Para os do SRA, coloquei para rodar utilizando bash e nohup. Ao checar na sexta-feira pela manhã, o subprograma classify do kraken ainda estava na lista do TOP. De tarde, no mesmo dia, o job já não estava mais lá. Bertolino comentou que provavelmente a equipe matou o job

propositadamente. Resolvi fazer um teste com uma amostra isolada, de 20 mb e com 12 threads em vez de 24 e chequei na segunda, dia 25/02 e o trabalho não estava na lista do top e não há sinal do output. Coloquei mais uma vez para anotar uma amostra isolada, de 15MGB: SRR6784993, utilizando bash e nohup e também submeti para fila. O job na fila permanece com o mesmo problema, já o que submeti no 'head node' ainda está listado no TOP por enquanto. Coloquei o trabalho no head node às 10:40.

8.10.3 Ferramenta microbiome no R

O professor recomendou utilizar a seguinte ferramenta: <https://microbiome.github.io/microbiome/#getting-started>. Talvez seja interessante para gerar as figuras de análises.

Lendo sobre o pacote, percebi que vou precisar que os dados estejam em um formato legível pelo pacote, o formato "phyloseq". Para transformar o dado, segue a página que explica como fazer isso <https://joey711.github.io/phyloseq/import-data.html>. A página recomenda baixar, entre os vários pacotes, o library(BiocManager).

Instalando ferramenta

Primeiro, deve-se ter R 3.5. Depois instalar o pacote devtools. Depois o pacote BiocManager.

Capítulo 9

Functional annotation of metagenomes

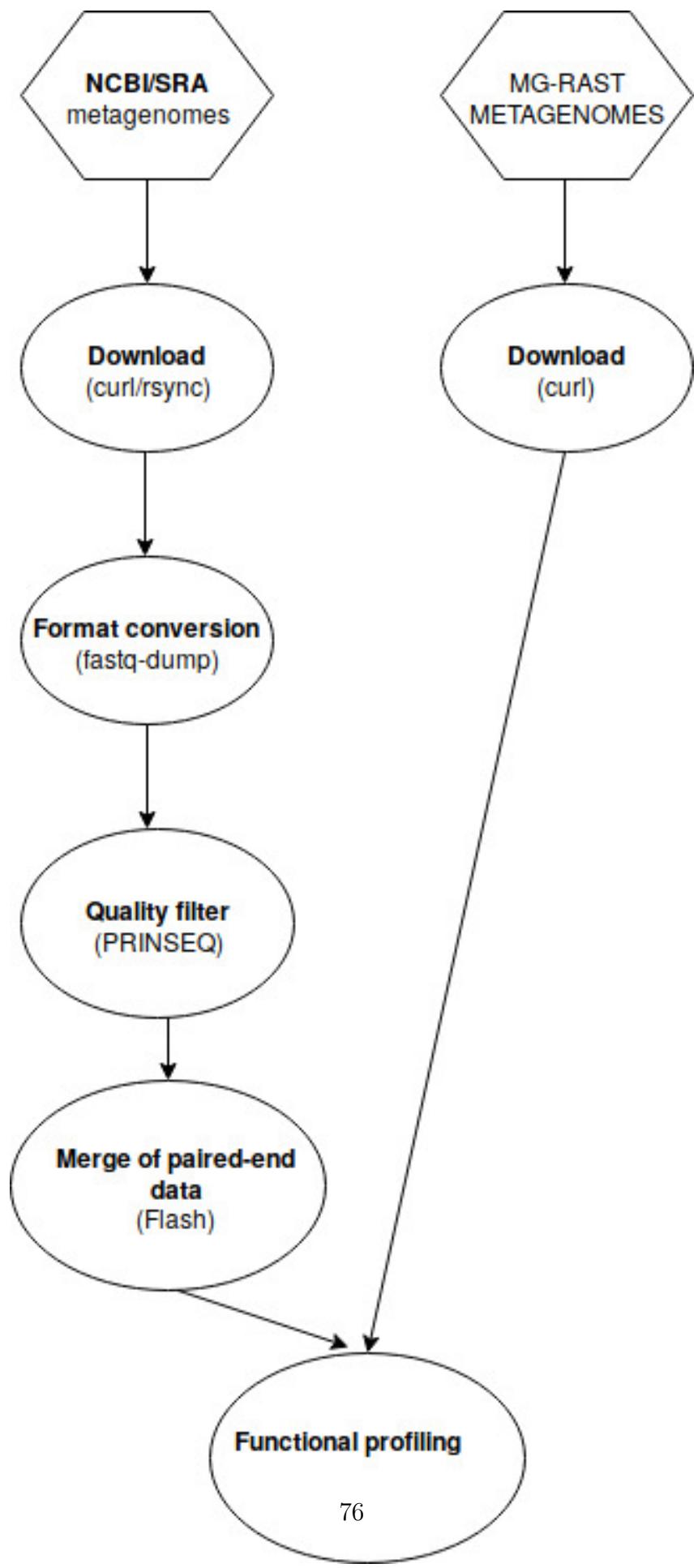


Figura 9.1: Pipeline of functional annotation

Capítulo 10

references

Articles list:

- 10.1371/journal.pone.0071301: Relata resultados que eu acreditava ter sido a primeira a encontrar
- 10.1038/nature14486: reconstruction of microorganism's genomes we use
- 10.1038/nm microbiol.2016.48: three of life, including the Candidate Phyla Radiation
- 10.1146/annurev.micro.57.030502.090759: speaks about the uncultured majority of microorganisms
- 10.1038/ismej.2016.174: revision of rare biosphere
- 10.1038/nrmicro3400: another revision of rare biosphere
- 10.1126/science.1224041: metabolic activities of *Candidatus Parcubacteria*, one of super-phyla of CPR
- 10.1128/MMBR.00009-08: Revision of bioinformatic methods and steps for metagenomic
- 10.1186/s40168-018-0428-1: Sponge as holobiont. Note: This article has a important information about microbial ecology: "Network and modeling analyses aim to disentangle the strength and nature (positive, negative, or neutral) of the interactions and predict their dynamics. Bacteria-bacteria network analysis of the core microbiota in different sponge species has revealed a low connective network with very few strong and many weak unidirectional interactions (i.e., amensalism [-/0] and commensalism [+/0] prevailed over cooperation [+/+] and competition [-/-]. These findings are consistent with mathematical models that predict that weak and non-cooperative interactions help to stabilize highly diverse microbial communities, whereas cooperation yields instability in the long term by fueling positive feedbacks"
- 10.1016/j.tim.2009.09.004: Microbial disease and the coral holobiont
- 10.3389/fmicb.2017.00618: Comparative Metagenomics of the Polymicrobial Black Band Disease of Corals

- 10.1038/nrmicro1643: The role of ecological theory in microbial ecology
- 10.1038/nrmicro3218: Explaining microbial genomic diversity in light of evolutionary ecology
- 10.1111/j.1462-2920.2009.01935.x: Metagenomic analysis of stressed coral holobionts
- 10.1038/nature06810: Functional metagenomic profiling of nine biomes
- 10.3389/fcimb.2014.00176: Microbes in the coral holobiont: partners through evolution, development, and ecological interactions
- 10.1038/ismej.2015.39: The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts
- 10.1111/j.1462-2920.2007.01383.x: Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*
- 10.1038/nm microbiol.2015.32: Metagenomics uncovers gaps in amplicon-based detection of microbial diversity
- 10.1038/ismej.2016.45: Challenges in microbial ecology: building predictive understanding of community function and dynamics
- 10.1111/j.1462-2920.2009.02113.x: Microbial functional structure of *Montastraea faveolata*, an important Caribbean reef-building coral, differs between healthy and yellow-band diseased colonies
- 10.1111/j.1758-2229.2010.00234.x:
- 10.1038/ismej.2011.116: Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges

Capítulo 11

Softwares, instalacao e linhas

Instalar o kraken-biome

- Folder: */home/leticia*
- Command: *pip install kraken-biom*
- Site: *https://github.com/smdabdoub/kraken-biom*

Para atualizacao do Lab-Book para o Git-Hub:

git commit

git push origin master

Para transferencia:

maquina remota para local: *scp leticia.cavalcante@login.sduumont.lncc.br:/scratch/ebiodiv/leticia.cavalcante@leticia/Documentos/dados*

11.1 Profiling metagenomes

Capítulo 12

Fundamentos teóricos e escrita do artigo

12.1 março de 2019

Decidi fundir os capítulos, após a conversa com a Amanda e compreender por alto por onde exatamente minhas leituras devem ir para fundamentar o meu trabalho, em vez de ficar lendo para acumular conhecimento sobre corais de forma não estruturada. A Amanda conversou comigo sobre fundamentar melhor as perguntas e resultados desejados para meu trabalho. Chegamos as seguintes conclusões:

Minhas perguntas são:

- 1) **Como (e se) as comunidades microbianas se diferenciam dado mudanças no estado de saúde, localidade, gênero, desenvolvimento e estudo?**
- 2) **Quais são as variáveis mais importantes para separação da comunidade microbiana?**
- 3) **A assinatura (biosfera rara) da comunidade também varia com os fatores acima?**

Hipótese teórica: sim, existe diferença na comunidade microbiana de corais dado diferença na saúde, localidade etc.

Hipótese operacional: a diferença se manifestará na estrutura(I), composição(II) e abundância(III)

Resultados

Para responder a pergunta 1, utilizaremos o nMDS, observando a estrutura(I) das comunidades e também para a pergunta 2, vendo qual a melhor condição que as diferencia. A abundância(elemento III) será um dos critérios para observarmos a diferenciação da comunidade e responder a pergunta 1, utilizando barras de abundância. Para responder a pergunta 3, utilizaremos diagrama de venn para cada condição, em que cada universo conterá filos únicos, compartilhando entre os diferentes estados das condições e aquilo que for compartilhado por todos será core. Sobre direcionar minha leitura de artigos, para que meu conhecimento seja mais estruturado, eu lerei artigos procurando respostas prováveis para minhas perguntas.

- 1) Leitura do artigo "Coral physiology and microbiome dynamics under combined warming and ocean acidification

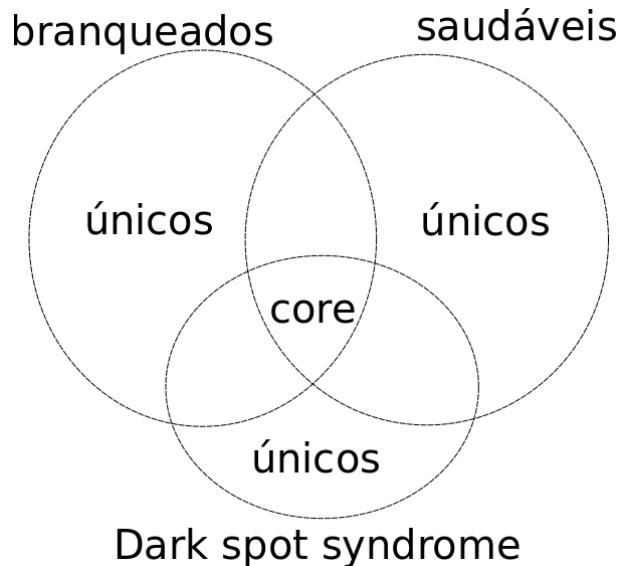


Figura 12.1: Diagrama de Venn exemplificando como eu visualizarei a assinatura da comunidade microbiana de corais em diferentes condições, utilizando a composição como elemento 2

Dessa forma, as figuras que devo fazer são:

- a) nMDS(s) em que eu colorirei as amostras utilizando as diferentes condições
- b) Com os nMDSs acima, verei qual é a condição que melhor separa as comunidades
- c) RF supervisionado pelas diferentes fatores que estou analisando
- d) Gráficos de abundância em que visualizo os filos que apareceram nos random forest(s)
- e) Diagramas de Venn

Uma das coisas que talvez eu deva discutir melhor é a parte de assinatura. A Amanda e eu conversamos que, muito provavelmente os filos mais abundantes não variem muito nas comunidades microbianas, mas a biosfera rara varie. Dessa forma, a biosfera rara será considerada como assinatura.

Capítulo 13

Meetings

Capítulo 14

Escrita do artigo

14.0.1 30 de outubro 2018

Na reuniao com Amaro, Miguel e o professor, mostrei os slides contendo os resultados obtidos com analises dos metagenomas utilizando a base de dados definitiva. Foram levantadas as seguintes questoes:

- As analises por familias não nos disseram muita coisa, alem de nao conter familias dos grupos nao cultivados por nao existir tal resolucao taxonomico. Por isso, utilizaremos apenas filo para trabalhar
- Fazer um PCA utilizando especies de coral sem identificar status de saude visualmente
- Fazer um PCA utilizando apenas amostras de corais saudaveis, identificando apenas genero

Tipo, se quero ver a diferenca existente entre as comunidades de generos de corais diferentes, devo fazer a analise só utilizando saudaveis em um caso e em outro sem identificar o estado de saude.

- Fazer um pca identificando apenas corais saudaveis e doentes, sem identificar visualmente o genero
- Fazer um pca utilizando apenas corais doentes
- utilizar mais filos indicados pelo random forest para fazer o pca
- PCA por especie
- Procurar por amostras de madracis que o professor trabalho - FEITO
- fazer uma planilha simplificada para Miguel e Amanda - FEITO (re-enviar a planilha, visto que encontrei as amostras e as incluirei na planilha de metagenomas)

14.1 30 de novembro

Visto que as figuras não ficarão prontas até o dia da reunião 02/12, eu conversei ontem com o professor sobre o que falar nessa reunião e ele pediu para que apresentássemos a estrutura proposta do artigo em vez de remarcar. Eu já tinha uma estrutura de tópicos do artigo pronta, arquivo: topic_paper.odt na pasta papers. Mostrei a Amanda, para saber se estava claro. Ela me ajudou a re-estrutura a parte dos results, com quais perguntas quero responder com as figuras que produzirei. Ficou da seguinte maneira:

1. Pergunta: Quais os filos mais abundantes e qual fator faz com que a abundância desses variem?
 - Fator saude: Existe diferença na abundância desses mais abundantes entre corais com diferentes estados de saúde?
Barplot, com X sendo corais saudáveis e doentes
 - Fator localidade: Existe diferença na abundância desses mais abundantes entre corais com diferentes localidades?
Boxplot, com X sendo diferentes localidades
 - Fator desenvolvimento: Existe diferença na abundância desses mais abundantes entre corais com diferentes tipos de desenvolvimento?
Barplot, com X sendo diferentes tipos de desenvolvimento
2. Pergunta: A estrutura da comunidade difere quanto a:
 - Estado de Saude?
 - nMDS de healthy vs diseased
 - Localidade?
 - nMDS de diferentes localidades
 - Desenvolvimento?
3. Pergunta: Qual a biosfera rara? Quais filos são ubíques? Tabela suplementar sugerida pela Amanda, em que as linhas são filos e as colunas será raridade a 1% e raridade a 5% e ubiquidade em 50% das amostras e 75%? Maneiras de visualizar?

Referências Bibliográficas