IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

Relatório do ‘Assignment 2’

High Throughput Sequencing at your fingertips

Realizado por:

Eduardo Palma nº201900054;

Marcos Torres nº201900019;

Nuno Melo nº201700465;

Ricardo Santos nº201700524;

Docente: Francisco Pina Martins

Unidade Curricular: Análise de Sequências Biológicas

Curso: Licenciatura em Bioinformática, 2º ano

Ano Letivo: 2020/2021

Instituição: Escola Superior de Tecnologia do Barreiro

Data: 25 de junho de 2021

Índice

Índice de Figuras

1. Introdução

Este Assignment 2 foi o último projeto do respetivo ano letivo 2020/2021 e para que conseguisse ser resolvido com sucesso teve de se utilizar diversos conhecimentos e tutoriais das aulas adquiridos ao longo do semestre.

O principal objetivo do problema era identificar as diferenças entre o cheiro das flores de um tipo de planta e para isso os exploradores procuraram por este tipo de planta, nas profundezas da floresta da amazónia. Entre estas plantas, alguns indivíduos tinham flores particularmente fedorentas (*Pungent*), enquanto a maioria dos indivíduos não tinha nenhum odor percetível (*Bland*). Alguns afirmaram que isso se devia ao facto de os indivíduos *Pungent* pertencerem a uma espécie diferente, que por acaso são parecidos com indivíduos *Bland*, enquanto outros afirmam que isso provavelmente se deve à variabilidade intraespecífica (ocorre dentro das espécies).

O sequenciamento de vários genes plastidiais e mitocondriais usando a tecnologia Sanger foi inconclusivo em relação ao problema acima, uma vez que o número de diferenças entre indivíduos *Pungent* e *Bland* era maior do que entre indivíduos *Bland*, mas não por qualquer margem significativa.

Como tal, os indivíduos *Pungent* e *Bland* foram sequenciados usando métodos RAD-Seq.

Para resolver este problema foi usado o método de *High Throughput Sequencing* (HTS). Há vinte anos, a tecnologia chamada HTS foi aplicada ao sequenciamento de cópias de DNA de comprimento parcial de "marcas de sequência expressa" para produzir uma visualização da biblioteca de genes expressa na amostra. Hoje, apesar de ser maior e de ter o mesmo objetivo sob o novo termo RNA-seq, que representa essencialmente os dados de sequência de genes que estão ativos numa determinada amostra biológica (Zhou et al., 2018).

2. Materiais e métodos

Para começar, foi necessário criar um ambiente no conda e ativá-lo, Figura 1.

```
conda create -n ipyrad
conda activate ipyrad
```

Figura 1 – Criação e ativação do ambiente ipyrad

Após alterar o ambiente no conda, clonou-se o repositório disponível em https://gitlab.com/StuntsPT/small_test_dataset/-/tree/master/Cornales e descomprimiu-se os ficheiros do tipo *fastq*, utilizando as ferramentas *xz-utils,* Figura 2.

```
unxz Cornales_R1.fastq.xz
unxz Cornales_R2.fastq.xz
```

Figura 2 – Descompressão dos ficheiros .xz

```
ipyrad -n cornales
```

Figura 3 – Criação do ficheiro dos parâmetros

Agora que se tem os parâmetros por *default*, alterou-se o que foi necessário, Figura 4. No caminho para os ficheiros *fastq,* usou-se "Cornales_R*" para que fossem lidos ambos os arquivos do Cornales. Inseriu-se também a localização dos *Cornales.barcodes,* Figura 5, o ficheiro completo está disponível no Assignment2_Anexos. Também se utilizou "*" nos *outputs_formats* para obter-se os ficheiros nos vários formatos pretendidos.

```
------- ipyrad params file (v.0.9.78)-------------------------------------------
cornales                       ## [0] [assembly_name]: Assembly name. Used to name
output directories for assembly steps
./ipyrad-assembly              ## [1] [project_dir]: Project dir (made in curdir
if not present)
./cornalesdata/Cornales_R*     ## [2] [raw_fastq_path]: Location of raw non-demul-
tiplexed fastq files
./cornalesdata/Cornales.barcodes   ## [3] [barcodes_path]: Location of barcodes
file
                               ## [4] [sorted_fastq_path]: Location of demulti-
plexed/sorted fastq files
denovo                         ## [5] [assembly_method]: Assembly method (denovo,
reference)
                               ## [6] [reference_sequence]: Location of reference
sequence file
ddrad                          ## [7] [datatype]: Datatype (see docs): rad, gbs,
ddrad, etc.
TGCAG,                         ## [8] [restriction_overhang]: Restriction overhang
(cut1,) or (cut1, cut2)
5                              ## [9] [max_low_qual_bases]: Max low quality base
calls (Q<20) in a read
33                             ## [10] [phred_Qscore_offset]: phred Q score offset
(33 is default and very standard)
6                              ## [11] [mindepth_statistical]: Min depth for sta-
tistical base calling
6                              ## [12] [mindepth_majrule]: Min depth for majority-
rule base calling
10000                          ## [13] [maxdepth]: Max cluster depth within sam-
ples
0.85                           ## [14] [clust_threshold]: Clustering threshold for
de novo assembly
0                              ## [15] [max_barcode_mismatch]: Max number of al-
lowable mismatches in barcodes
2                              ## [16] [filter_adapters]: Filter for adapters/pri-
mers (1 or 2=stricter)
35                             ## [17] [filter_min_trim_len]: Min length of reads
after adapter trim
2                              ## [18] [max_alleles_consens]: Max alleles per site
in consensus sequences
0.05                           ## [19] [max_Ns_consens]: Max N's (uncalled bases)
in consensus
0.05                           ## [20] [max_Hs_consens]: Max Hs (heterozygotes) in
consensus
4                              ## [21] [min_samples_locus]: Min # samples per lo-
cus for output
0.2                            ## [22] [max_SNPs_locus]: Max # SNPs per locus
8                              ## [23] [max_Indels_locus]: Max # of indels per lo-
cus
0.5                            ## [24] [max_shared_Hs_locus]: Max # heterozygous
sites per locus
0, 90, 0, 0                    ## [25] [trim_reads]: Trim raw read edges (R1>,
<R1, R2>, <R2) (see docs)
0, 0, 0, 0                     ## [26] [trim_loci]: Trim locus edges (see docs)
(R1>, <R1, R2>, <R2)
*                              ## [27] [output_formats]: Output formats (see docs)
                               ## [28] [pop_assign_file]: Path to population as-
signment file
                               ## [29] [reference_as_filter]: Reads mapped to this
reference are removed in step 3
```

Figura 4 – Ficheiro dos parâmetros

```
Bland-26        CACCA
Bland-14        TCTTGG
Pungent-5       AAGACGCT
Bland-50        CAGAGGT
Bland-58        GAAGCA
Bland-52        ACGCGCG
Bland-41        ACCAGGA
Bland-6          ATAGAT
Bland-43        TCACGGAAG
```

Figura 5 – Exemplo de alguns barcodes

Utilizou-se o ficheiro dos parâmetros no *ipyrad* nos 7 passos utilizados para a análise, mostra-se na figura 6 como exemplo o primeiro passo. No primeiro passo verificou-se quais os indivíduos que não tinham informação biológica associada e removeu-se os *barcodes* dos mesmos. Os *barcodes* removidos foram Bland-22, Bland-28, Bland-51, Bland-70, Bland-82 e Bland-90.

```
ipyrad -p params-cornales.txt -s 1 -c 3
```

Figura 6 – Ipyrad

Por fim, para perceber se os indivíduos são ou não de espécies diferentes, utilizou-se árvores filogenéticas para visualizar e comparar as diferenças entre as espécies.

O programa utilizado para produzir a árvore de *Maximum Likelihood* foi o RAxML (raxmlHPC v8.2.12) (Stamatakis, 2014), Figura 7, e para visualizar o *FigTree (*FigTree v1.4.4).

```
raxmlHPC-PTHREADS-AVX -f a -s cornales.phy -m GTRGAMMA -x 2525 -p 2525 -N 100 -n Cornales_ML
```

Figura 7 – Criação da árvore de Maximum Likelihood

Para fazermos uma comparação teve-se de criar uma segunda árvore utilizando o método de *MrBayes* (mb v3.2.5) (Ronquist et al., 2011), que permitiu diminuir o tempo de processamento.

Além das árvores, também foi realizada uma Análise de Componentes Principais, usando o *RStudio* e dois pacotes provenientes do repositório *Bioconductor,* o pacote *pcaMethods* e o pacote LEA.

Para esta análise foi utilizado o ficheiro *geno* obtido a partir do *ipyrad*.

```
corn_data = read.geno("cornales.geno")
```

Figura 8 - Carregamento do ficheiro geno

Esta análise foi realizada fundamentalmente para comparar e confirmar os resultados obtidos nas árvores.

```
corn_pca=pcaMethods::pca(corn_data, scale="none", center=T, nPcs = 2, method="nipals")
slplot(corn_pca,
       scol=cores,
       scoresLoadings=c(TRUE, FALSE),
       sl=NULL,
       spch="x")


legend("bottomright",
       legend=(unique(types)),
       col=unique(cores),
       pch="x",
       title="Plant Types")
```

Figura 9 - Criação do Score Plot das Componentes Principais

3. Resultados

Começando com a árvore obtida pelo RAxML, pode-se facilmente identificar as duas espécies, os *Bland e os Pungent,* a árvore está suportada com um valor de *bootstrap* de 96%, Figura 8.

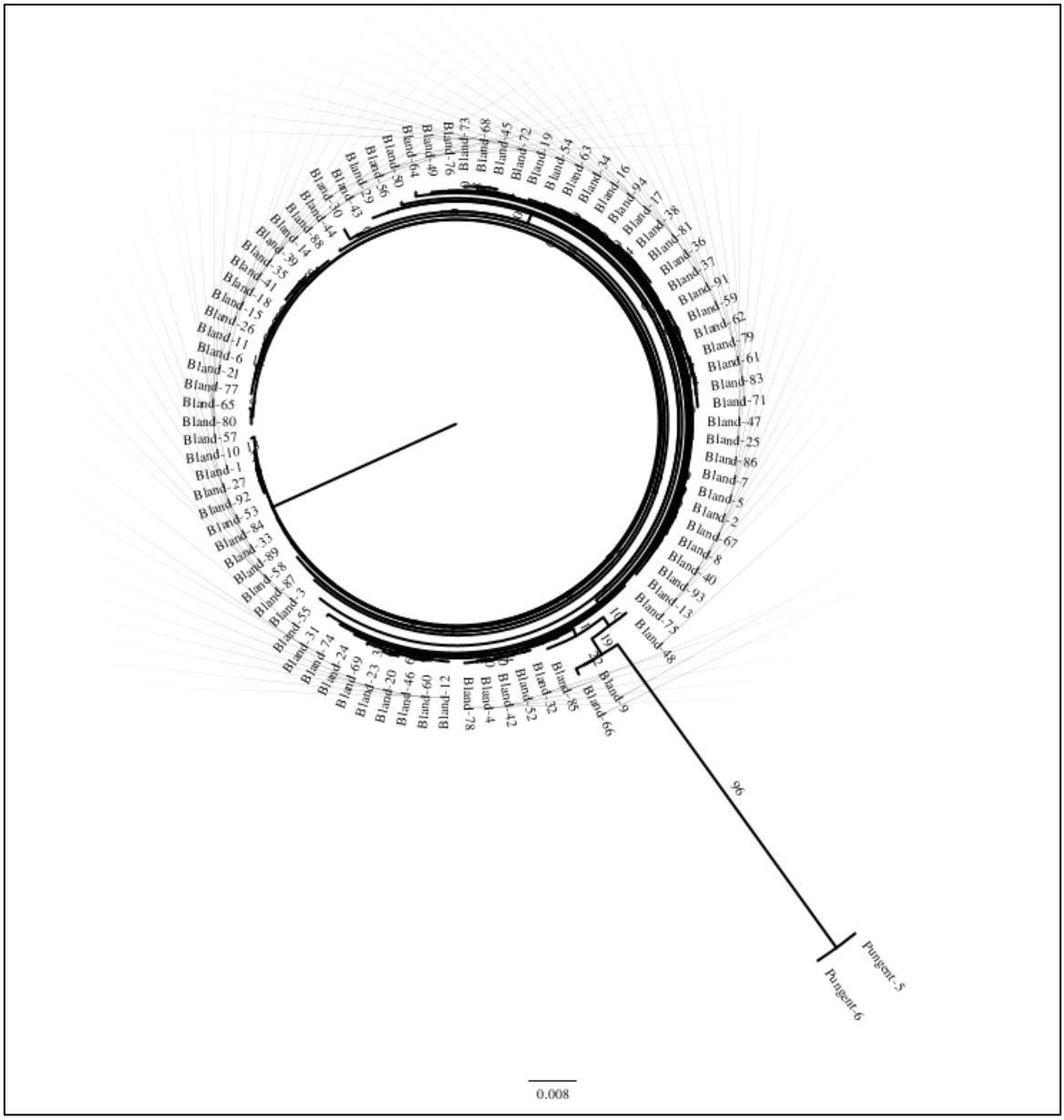

Figura 10 – Árvore de ML (Maximum Likelihood) unrooted com bootstrap 96%.

De seguida, utilizando o *MrBayes* construiu-se outra árvore para conseguir tirar mais conclusões e comparar com a árvore de *Maximum Likelihood* construída anteriormente, na Figura 12.

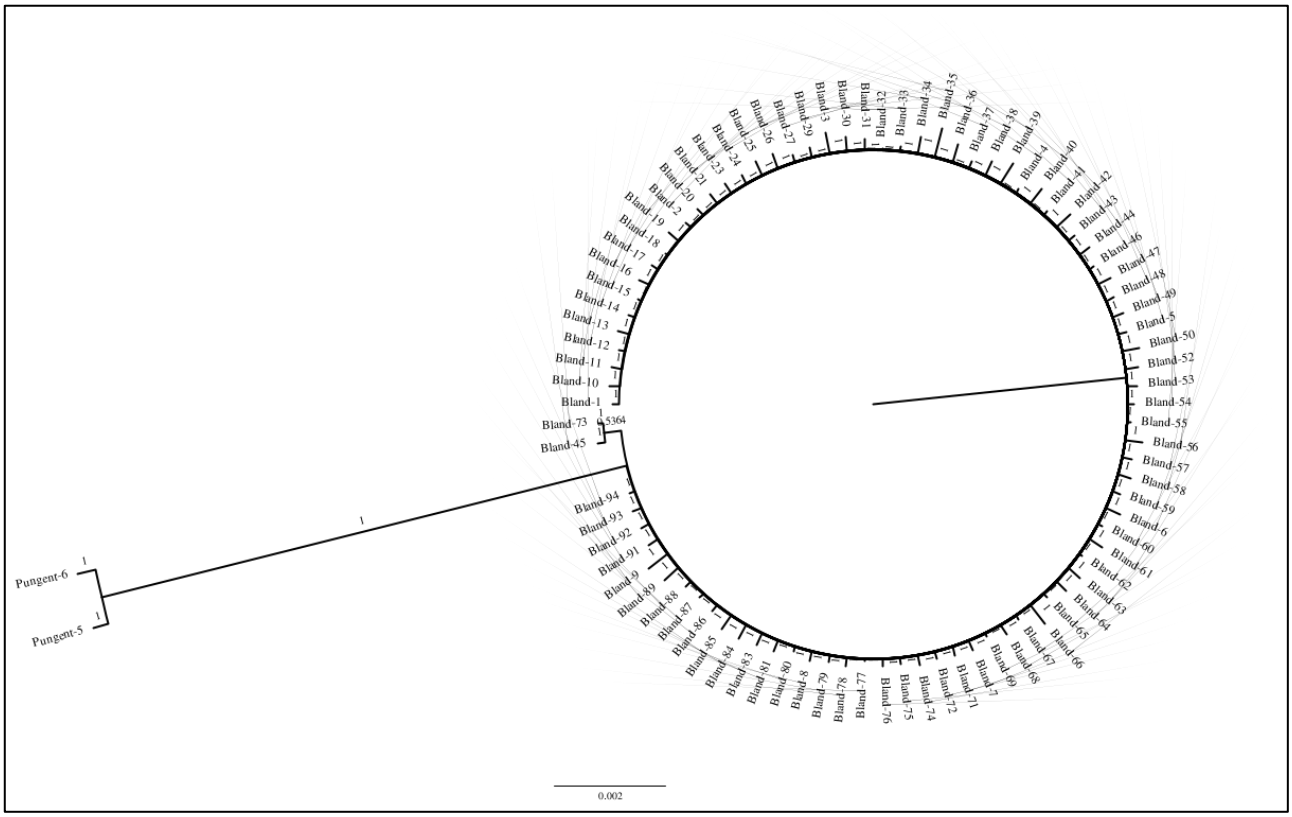

Figura 11 – Árvore de BI

Com as duas árvores construídas consegue-se perceber que o valor de bootstrap de 96% da árvore de ML e de *posterior probability* de 100% da árvore de BI são muito idênticos, o que se dá a entender que os indivíduos *Bland* e *Pungent* estão em grupos diferentes.

3.1 Análise de Componentes Principais em R

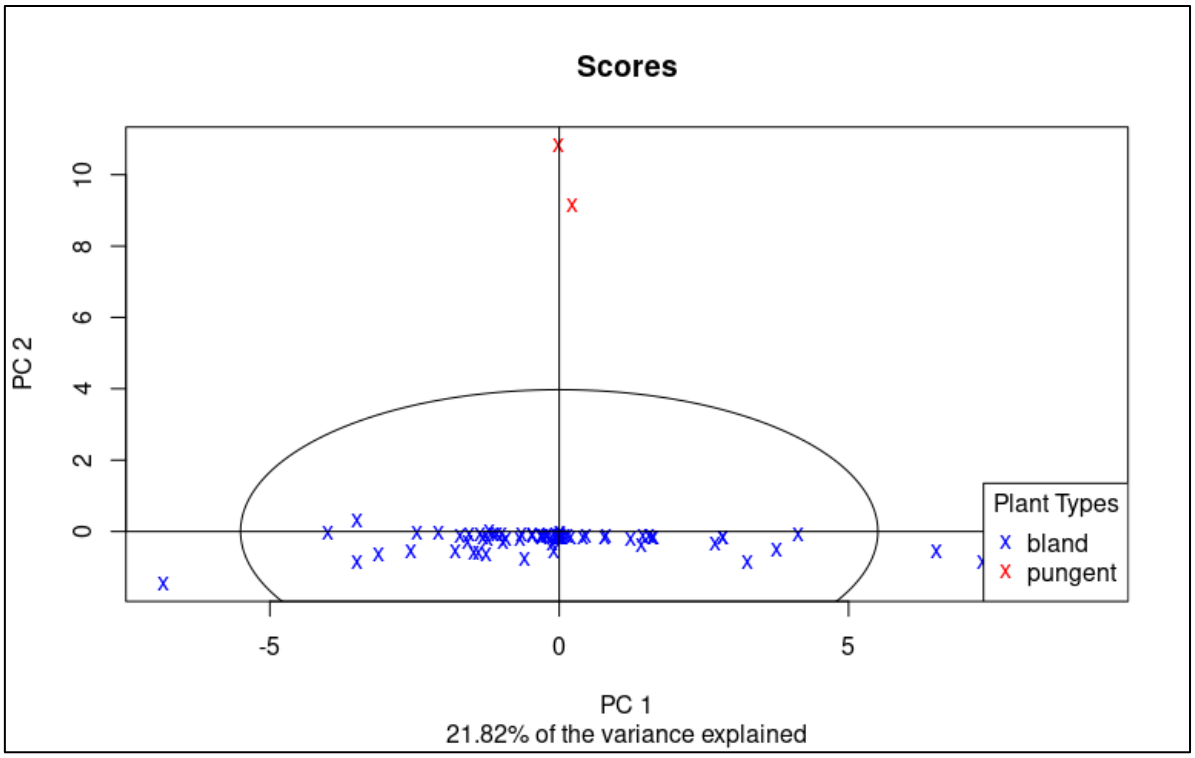

Figura 12 – Score Plot das Componentes Principais

Na Figura 12, apesar da presença de alguns *outliers,* cuja variação não é explicada pelas componentes, observa-se que não só o número de indivíduos *Bland* é superior ao de indivíduos *Pungent*, como já observado nas árvores, estes também estão maioritariamente agrupados entre si, em comparação com os indivíduos *Pungent,* que possuem uma variação maior.

4. Discussão

Por fim, os resultados sugerem uma clara diferença entre as duas espécies, *Bland* e *Pungent,* e com isto prova que existe diferenças entre a informação genética de cada um dos indivíduos das diferentes espécies. Daqui em diante, os analistas poderão realizar os seus estudos suportando-se nestes resultados, tendo uma melhor perceção das diferenças existentes na filogenética destas duas espécies.

5. Bibliografia

Ronquist, F., Huelsenbeck, J., & Teslenko, M. (2011). MrBayes Version 3.2 Manual: Tutorials and Model Summaries. *Manual MrBayes*, *July*, 1–103.

Stamatakis, A. (2014). RAxML. *Manual/Tutorial*, *1*, 1–5.

Zhou, W., Ji, X., Obata, S., Pais, A., Dong, Y., Peet, R., & Xiang, Q. Y. (Jenny). (2018). Resolving relationships and phylogeographic history of the Nyssa sylvatica complex using data from RAD-seq and species distribution modeling. *Molecular Phylogenetics and Evolution*, *126*, 1–16. https://doi.org/10.1016/j.ympev.2018.04.001