Assignment 2

High Throughput Sequencing at your fingertips

Instituto Politécnico de Setúbal – Escola Superior de Tecnologia do Barreiro

Análise de Sequências Biológicas

julho 2022

Prof. Francisco Pina Martins

2021/2022 Bioinformática

Alexandre Duarte n. °202000198

Diogo Cabrita n. °202000212

Índice

Índice de Figuras

Introdução

No âmbito da UC de Análise de Sequências Biológicas, foi proposto pelo docente a realização deste relatório afim de serem utilizadas e aprimoradas as habilidades e capacidades de *High Throughput Sequencing*, tendo por base os métodos e os conteúdos lecionados nas aulas, durante o semestre.

High Throughput Sequencing, ou sequenciamento de alto rendimento, é usado e aplicado para fornecer uma abordagem de seleção dos genomas com maior rigor determinando assim, a ordem das bases num fragmento de DNA. Ao evidenciar essa ordem, é possível efetuar um estudo sobre os genes, as regiões reguladoras, os RNAs, as proteínas de um organismo e até comparar sequências para o desenvolvimento e conhecimento científico.

Foram considerados vários métodos que, após uma análise reflexiva, se considerou o método A*dmixture Plot*, como o mais proficiente.
Esta ferramenta é muito utilizada/vulgarizada para analisar dados SNP. A forma de trabalho é com *clustering* não supervisionado de grandes números de amostras o permite, a que cada indivíduo esteja presente nos *clusters*. (Schiffel, Peltzer, & Clayton, 2016)

Da mesma forma, para a concretização deste trabalho, será utilizada outra ferramenta, denominada de PCA (*Principal Component Analisys*), com o intuito de ser aplicada a redução de dimensionalidade. Neste campo, pretende-se identificar as dimensões que melhor diferenciam o conjunto de dados em análise, ou seja, os seus componentes principais. (Lima, Tapadas, & Masson, 2018)

O caso em estudo é a distribuição da espécie *Quercus suber,* localizada em diferentes espaços territoriais, para que seja possível avaliar a história evolutiva da espécie, do ponto de vista de DNA nuclear, aprimorar os *insights* sobre como a adaptação local está moldada ou adaptada as bases genéticas da espécie e para prever como a espécie consegue responder às mudanças climáticas de uma perspetiva genética.

A história evolutiva do sobreiro (*Quercus Suber*) foi estudada anteriormente, através da utilização de múltiplas metodologias, em diferentes faixas geográficas.
Observando o mapa abaixo [FIG.1], os autores (Martins, Batista, Pappas, & Paulo, 2018) apresentam um mapa com a distribuição geográfica do sobreiro à volta do mediterrâneo, sendo que as áreas sombreadas a verde representam o alcance da espécie e as localidades das amostras representam os locais de onde foram recolhidas.

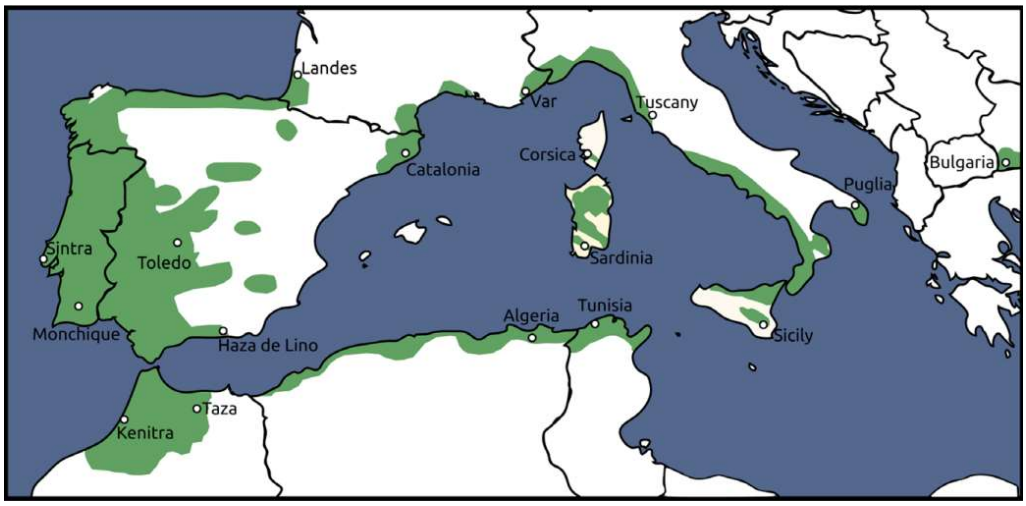

Figura 1-Mapa geográfico da distribuição das amostras em estudo0

Em resultado de estudos recentes, sobre esta temática, sugere-se que o sobreiro é dividido em quatro linhagens estritamente definidas.
A 1.ª e 2.ª linhagem variam do sudeste da França a Marrocos, incluindo a Península Ibérica e a Ilhas Baleares, a 3.ª linhagem vai desde a região de Mónaco até *Algeria* e Tunísia, incluindo as ilhas da *Corsica* e da *Sardinia*. A 4.ª linhagem abrange toda a península Itálica, incluindo Sicília.

Com base em estudos em marcadores plastidiais destas linhagens, foi demonstrado que dificilmente estes indivíduos compartilham os mesmos haplótipos, ou seja, não existe uma correspondência direta entre o grupo de alelos e o *locis* adjacentes.

No entanto, através estudos posteriores, baseados em DNA nuclear, sugerem um cenário diferente, demonstrando que a espécie não é tão estritamente dividida ou diferente como aparenta. Desta forma, pode-se concluir que esta diferenciabilidade de segregação só está presente nos marcadores plastidiais.

Problema Biológico Original

O Sobreiro ou *Quercus Suber* é uma árvore maioritariamente presente em lugares como a Península Ibérica, o norte de África, Itália e partes da França sendo extremamente bem-adaptada ao clima mediterrânico, devido às suas raízes profundas que possibilitam captar água em grande profundidade. Da mesma forma, esta espécie apresenta folhas com cutículas que ajudam a impedir o excesso de transpiração, fazendo com que a perda de água seja menor, como também um material espesso e esponjo que a protege dos incêndios – a cortiça. (Tiago & Rosário, 2019)

Este tecido florestal, produzido por esta espécie, tem vindo a desempenhar um papel significativo a nível económico e social, ao longo dos tempos. O desenvolvimento tecnológico também induziu à maior versatilidade dos produtos originados. Outrora destinado à produção de rolhas para garrafas, atualmente verifica-se a comercialização para o fabrico de isolantes térmicos, tecidos de cortiça, cordéis, na indústria aeronáutica e na indústria da música, sendo considerada como um excelente isolante sonoro. (Tiago & Rosário, 2019)

Com a elevada expansão da população humana e as tecnologias emergentes e todos as novas formas de transporte e indústria criadas, fez com que dezenas de gases, toxinas e combustíveis fossem emitidos para a atmosfera e para o solo, contribuindo para o aumento do efeito de estufa e diminuição da camada de ozono.

Presentemente a população global está a viver num período caraterizado pelo aquecimento global, onde as temperaturas médias do mar, do solo e da atmosfera estão a aumentar progressivamente, fazendo com que muitas espécies se tenham extinguindo, ao longo deste tempo de alterações climáticas, e outras em risco de extinção.

O aumento populacional, o consumo desacerbado, a evolução tecnológica, entre outras, tem vindo a registar indícios de uma passível destruição ambiental e de espécies animais e vegetais. No entanto, em resultado deste comportamento e, como autodefesa, muitas espécies têm vindo a se adaptar a estas alterações.

No decorrer do *paper,* os autores (Martins, Batista, Pappas, & Paulo, 2018) apresentaram vários problemas e questões que são colocadas para se conseguir avaliar a história evolutiva do *Quercus suber* sabendo, à partida, que cada amostra provém de um lugar distinto.

O estudo tenta demonstrar sinais de seleção natural como também possíveis relações entre as espécies para se perceber de que forma a adaptação local está moldada/adaptada às bases genéticas da espécie. A abordagem relativa às alterações climáticas, apresentando previsões sobre a evolução da espécie em estudo, é uma contante ao longo do estudo.

Materiais e Métodos do Paper Original

Os autores do *paper* (Martins, Batista, Pappas, & Paulo, 2018) decidiram fazer a divisão dos métodos em 5 etapas.

Etapa 1 - Amostra e recolha de dados ambientais

Iniciando a primeira etapa, realizou-se uma recolha de dados com a finalidade de fornecer uma visão geral das amostras em estudo.

Foram recolhidas 17 amostras de locais distintos abrangendo a maior parte da distribuição da nossa espécie *Quercus Suber*.

No decorrer das amostras foram recolhidas folhas frescas das diferentes espécies nos seguintes locais, a saber: 6 da Bulgária, da Corsica, da Kenitra, de Monchique, da Puglia, da Sardinia, da Sicília, da Tuscany, da Tunísia e de Var (60 amostras) e 5 espécies da Algeria, da Catalonia, da Haza de Lino, de Landes, de Sintra, da Taza e de Toledo (35 amostras) perfazendo um total de 95 amostras de 17 indivíduos.

Com a observação geográfica da espécie através da Fig.1 e com outras pesquisas efetuadas pelos autores (Martins, Batista, Pappas, & Paulo, 2018), chegaram à conclusão que os carvalhos do tipo *Quercus Suber* na Bulgária não são de origem natural, mas sim de origem artificial, resultado da intervenção humana.

De forma a ser mais fácil a análise e observação das coordenadas e o número de amostras relacionadas com uma determinada região, foi criada uma tabela – Tabela 1 com os dados recolhidos.

Sample site	Latitude (decimal deg.)	Longitude (decimal deg.)	Number of sampled individuals
Algeria	36.5400	7.1500	5
Bulgaria	41.43	23.17	6
Catalonia	41.8500	2.5333	5
Corsica	41.6167	8.9667	6
Haza de Lino	36.8333	−3.3000	5
Kenitra	34.0833	−6.5833	6
Landes	43.7500	−1.3333	5
Monchique	37.3167	−8.5667	6
Puglia	40.5667	17.6667	6
Sardinia	39.0833	8.8500	6
Sicilia	37.1167	14.5000	6
Sintra	38.7500	−9.4167	5
Taza	34.2000	−4.2500	5
Toledo	39.3667	−5.3500	5
Tunisia	36.9500	8.8500	6
Tuscany	42.4167	11.9500	6
Var	43.1333	6.2500	6
Total	—	—	95

Tabela 1 - Coordenadas e número de amostras em estudo

Etapa 2 - Preparação da biblioteca e dos sequenciamentos

Nesta etapa, os autores (Martins, Batista, Pappas, & Paulo, 2018) decidiram recolher o DNA, das amostras em estudo – folhas frescas-através de um procedimento que submetia as folhas moídas em nitrogênio líquido.

A quantidade total de DNA extraído foi quantificado através de estudo em espectrofotometria utilizando a ferramenta *Nanodrop* e a sua integridade foi verificada em gel de Agarose.
As amostras de DNA foram então diluídas numa concentração para serem criados processos de forma a determinar diferenças ao nível da composição genética das espécies em análise.

O sequenciamento foi feito numa única célula de fluxo *Illumina HiSeq,* através de uma enzima de corte pequeno (*EcoT22I*), devido ao grande tamanho do genoma do *Quercus Suber* ficando mais fácil de se proceder à analisada com este tipo de corte.

Etapa 3 - Análises dos dados de genomas

Nesta etapa, os autores (Martins, Batista, Pappas, & Paulo, 2018) recolheram um conjunto de dados brutos de GBS (*Genotyping by sequencing*) para depois serem analisados, utilizando o programa *IPYRAD* v0.7.24, através do ambiente anaconda com os programas, *MUSCLE* v3.8.31 e o *VSEARCH* v2.7.0.

Foi criada uma nova montagem das sequências, mas, para que fosse garantida as leituras de mtDNA (*mithocondrial DNA*) e cpDNA (*chloroplast DNA*), estas foram filtradas através de análises *downstream* onde foram "chamados" através do comando *denovo-reference* do programa *IPYRAD*.

As análises *downstream* foram feitas através do programa *GNU Make*. O propósito destas análises foi de não permitir que o processo fosse tratado como um objeto fechado ou seja, tentaram fornecer um ambiente completo para ser reproduzo, estudado e modificado por toda a comunidade científica.

Desta forma, os genomas mitocondriais filtrados foram: o *Populus davidiana* (KY216145.1), *Pyrus pyrifolia* (KY563267.1) e *Rosa chinensis* (CM009589.1). Os genomas cloroplastidiais filtrados foram: *Quercus rubra* (JX970937.1), *Quercus aliena* (KU240007.1) e *Quercus variabilis* (KU240009.1).
Foi criado um certo tipo de parâmetros incluídos no GBS para o tipo de dados, onde o limite de *clustering* tem valor de 0.85, o *mindepth* de 8, com uma incompatibilidade máxima de código de barras de 0.

Após este processo, os autores (Martins, Batista, Pappas, & Paulo, 2018) concluíram que para a viabilidade deste estudo seria necessário recolher em cada localidade uma amostra que

tinha que representar pelo menos de três indivíduos, para que um dos SNP fosse chamado. Apenas as localidades de *Kenitra* e *Taza* foram excluídas, onde foi apenas necessário um indivíduo devido à abrangência desta espécie nesses locais.

Todos os arquivos *fastq* foram subjugados a arquivos de leitura de sequências do site NCBI (*database SRA*) que fornecem informações biológicas, como *BioProject* de ACCN - PRJNA413625.

Todos estes dados processados, através do programa *IPYRAD,* foram retidos através do programa *VCFtools* v0.1.14, onde foram criados critérios que impunham que cada amostra tinha de estar representada em pelo menos 40% dos SNPs e, após esse processo, cada SNP em pelo menos 80% dos indivíduos.

As conversões de arquivos foram realizadas através do programa *PGDSpider* v2.1.0.0, exceto para os ficherios *BayPass* e formatos *SelEstim*, onde os scripts *geste2baypass.py* com *commit* "b99636e" e *gest2selestim.sh* com *commit* de "f74f66b" foram recolhidos pois estas versões utilizadas neste programa não conseguem ser lidas com nenhum desses formatos de ficheiros.

Foram também utilizados testes de estatísticas descritivas, usando o programa *Genepop* v4.6, de forma a ser possível determinar um eventual efeito de isolamento de alguma destas espécies. Com isto, os autores (Martins, Batista, Pappas, & Paulo, 2018) excluíram os indivíduos de amostras provenientes da Bulgária devido à sua origem artificial, já referido anteriormente.

Etapa 4 - Deteção de *outliers* e associações ambientais

Esta etapa teve como objetivo recolher apenas os SNPs que foram indicados como *outliers* por ambos os programas, uma vez que são considerados forasteiros para o propósito do estudo. Estes *outliers* pioram a análise e induzem a apresentação de valores anormais na amostra.

Foi possível então, fazer a deteção destes valores através de dois programas, *SELESTIM* v1.1.4 e *BAYESCAN* v2.1, pois estes dois programas tem uma função que possibilita uma menor taxa de falsos positivos ajudando, assim, na identificação dos valores pretendidos.
Os falsos positivos são um dos grandes problemas na realização deste tipo de análise devido aos problemas e erros possíveis que estes podem produzir.

Depois da deteção dos *outliers* , os autores (Martins, Batista, Pappas, & Paulo, 2018) fizeram a criação das associações utilizando um fator de *Bayes* (BF) acima de 15, onde apenas estes são considerados significativos. As restantes análises de associação foram realizadas excluindo as amostras das espécies provenientes da Bulgária pelas mesmas razões explicadas anteriormente.

Com esta análise, as sequências que continham *loci outliers*, ou SNPs associados a uma variável ambiental, foram consultadas comparando com o genoma de Q. *Lobata* v1.0 usando *BLAST v2.2.28* com o valor de e-valor de 0,00001 de forma a serem encontradas as associações ambientais dentro da espécie.

Etapa 5 - Estrutura da população

Aqui usaram dois métodos distintos para agrupar os indivíduos, de forma a entender o padrão geral de apenas um indivíduo ou do total de indivíduos, através da análise de componentes principais (PCA) como também de *MAVERICK*.

O PCA foi realizado através do *script snp_pca_static.R* a partir do *commit* "bb2fc45".
Para interpretar corretamente os resultados das análises destes agrupamentos é importante estimar o valor de K, que representa quantos *demes* em que os dados podem ser agrupados.

O *software MAVERICK* foi utilizado principalmente pela sua função de poder fazer estimativas de *clusters*, devido à sua inovação para estimar o melhor valor de K, que tem demonstrado um grande desempenho e avanço nestas questões, comparando a outros métodos que já tinham sido utilizados.

O *MAVERICK* foi subjacente ao programa *Structure_threader* v 1.2.2 que foi executado para valores de K entre 1 e 8, para ser escolhido o valor mais adequado de K.
Estas metodologias são usadas para se observar algumas indicações, nomeadamente se a adaptação local é responsável pelo padrão observado, ou não.

Conclusões do Paper Original

Neste *paper* (Martins, Batista, Pappas, & Paulo, 2018) foram recolhidas amostras de indivíduos de *Quercus Suber* com a finalidade de se avaliar a estrutura populacional e os padrões genéticos presentes, como também o impacto das alterações climáticas e a adaptação das espécies.

Inicialmente houve uma redução do genoma da espécie *Querceus Suber*, devido ao seu grande tamanho original, para se descobrir os *SNPs* da mesma.

Após rigorosa filtragem, foi criado um conjunto de 1.996 *SNPs* com a finalidade de serem usados para este estudo. Perante os dados, foram recolhidas amostras dos diferentes tipos de espécies de sobreiro *Quercus Suber* em dezassete locais distintos localizados a norte e a sul do mar mediterrâneo.
Das noventa e cinco amostras recolhidas, e numa primeira análise, foram descartadas doze amostras devido à sua baixa quantidade de sequências, durante o processo de recolha, resultando na retenção de oitenta e três indivíduos. De seguida, foram retiradas mais duas sequências ficando com o valor final de oitenta e um indivíduos.

Reitera-se que os autores (Martins, Batista, Pappas, & Paulo, 2018) decidiram não filtrar a amostra da Bulgária por não ser uma espécie local, mas sim proveniente de uma plantação artificial.
As amostras recolhidas para o estudo foram folhas frescas das árvores, em que o DNA é extraído através dessas folhas moídas com nitrogénio líquido para, mais tarde, ser possível comparar cada espécie e tentar perceber até que ponto, todas estas amostras, independentemente do seu local de origem, estão relacionadas geneticamente.
Com isto, foram feitos testes e experiências de forma a perceber-se como estas espécies se defendem contra as alterações climáticas e até que ponto estão aptas para se conseguirem defender contra as variações do clima.

A espécie em estudo tem muitas formas de se adaptar e, por vezes, evitar a sua extinção, conseguindo responder às mudanças a que está sujeita, nomeadamente, alterando a sua área geográfica ou adaptando-se às novas condições ambientais.

Neste caso, como se trata de uma espécie de árvore, o sobreiro consegue-se adequar de uma forma muito mais fácil do que a maior parte dos seres vivos existentes no planeta, pois as árvores não necessitam de tantas fontes de energia e de recursos como outros seres vivos necessita.

Os recursos energéticos, como por exemplo a água e o sol, são essenciais à sobrevivência das espécies que recorrerem a estes recursos de uma forma distinta. Existem seres vivos mais

dependentes destes recursos do que outros, os mais sujeitos acabam por não conseguir sobreviver.
No caso específico das árvores, e na maioria dos casos, estas espécies adaptam-se aos constrangimentos gerados em locais que estão em constante alteração, conseguindo atingir uma longevidade muito significativa.

Resumindo, os autores (Martins, Batista, Pappas, & Paulo, 2018) chegaram à conclusão, tendo por base as variáveis consideradas e os locais de amostra, que o sobreiro não exige grandes mudanças nas frequências alélicas para sobreviver às previstas alterações climáticas, ou seja, esta espécie conseguiu não ser afetada ao longo do seu tempo de existência.

Objetivos

Depois da reflexão, análise e leitura do *paper original,* este projeto pretende comparar os resultados obtidos pelos autores (Martins, Batista, Pappas, & Paulo, 2018) com os alcançados com o nosso trabalho, onde se vai tentar chegar às mesmas conclusões ou a outras de igual forma relevantes.

Começou tudo com a recolha de dados através da base de dados SRA do NCBI onde se conseguiu obter as 95 sequências a partir da ACCN PRJNA413625 (Martins, Quercus suber, 2017). Depois deste processo foi feita a criação dos ficheiros fastq para, com o programa *ipyrad*, conseguir-se os dados e ficheiros precisos para continuar com a análise da espécie.

Como principal objetivo e aplicando os dois métodos anteriormente referidos, denominadas de PCA (*Principal Component Analisys*), foram identificados os componentes principais para o conjunto de dados e o *admixture plot* como ferramenta para analisar os *SNPs* e conseguir tirar análises, através dos diferentes gráficos criados.

Materiais e Métodos utilizados

Recolha de dados das amostras e extração

Como referido nos objetivos começou-se pela extração das sequências e para isso foi preciso, primeiramente, vários dados das amostras como por exemplo os *ids*, nome, *run accession* e localização geográfica. Para isso, foi utilizado um dos pacotes dos *E-Utilities* [FIG.2], o *efetch*, como ferramenta para consultar e retirar informações de ficheiros XML do NCBI.

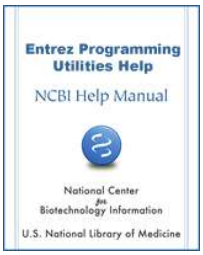

Figura 2 - Entrez Programming Utilities Help

Foi usado o conda [FIG.3] para criação de ambientes (*base, sra, ipyrad, structure*) e obter vários programas tais como o *sra-tools* (fasterq-dump), *ipyrad*, *vcftools* e *structure_threader*, para a execução dos *scripts* e análise das sequências.

Figura 3 - CONDA

Usámos, então o programa *Sra-tools* instalado através do conda para ser possível usar um comando (*fasterq-dump*) que extrairá as sequências SRA (*Sequence Read Archive*) da base de dados NCBI em formato *fastq*.
Foi usada também, a ferramenta *gzip* [FIG.4], que serviu para nada mais nada menos do que comprimir os ficheiros necessários, nomeadamente os *fastqs*, para depois serem usados nos outros programas e as restantes análises serem realizadas.

Figura 4 - Gzip

Análise dos dados do genoma

Para a análise das sequências do *Quercus suber* usou-se o programa *ipyrad* v.0.9.84 [FIG.5] instalado através do conda como o *sra_tools*, para a transformação dos dados brutos e para a recolha de ficheiros que podem ser usados para análise *downstream*.

Na execução deste programa 12 sequências foram descartadas devido à pouca quantidade de expressão das mesmas e assim ficando só com um total de 83 amostras.
Através deles conseguiu-se identificar o número total de *loci* que neste caso é 13147 e consegue-se retirar a matriz estatística dos alinhamentos onde o tamanho da matriz de sequências tem 70,28% *missing sites* e o tamanho da matriz de SNPs tem 71,09% *missing sites*.

Figura 5 - Ipyrad

Estruturação da população e gráficos

Relativamente à estruturação da população usou-se a linguagem e o programa *python* v3.7.12 [FIG.7], para a execução do ficheiro vcf.parser.py que foi preciso para a criação de um ficheiro requerido para o programa *structure_threader* [FIG.6] e assim obter o *admixture plot*.

Figura 6 - Structure_threader

Figura 7 - Python

Para a criação dos gráficos do PCA foi preciso alinhar o ficheiro. indfile usado para o programa *structure_threader* e depois feita a fusão do *file* alinhado com o ficheiro *assignment2.str* dado como um dos outputs do *ipyrad*.
Tendo os nossos dados prontos, o próximo passo foi realizado no RStudio v4.1.1 [FIG.8], onde através das suas ferramentas se conseguiu criar os *plots* PCA através do *script pca.R*.

Figura 8 - Rstudio

Os *scripts* usados para a realização das análises descritas acima nos materiais e métodos utilizados estão disponivíes num repositório no github.

Resultados

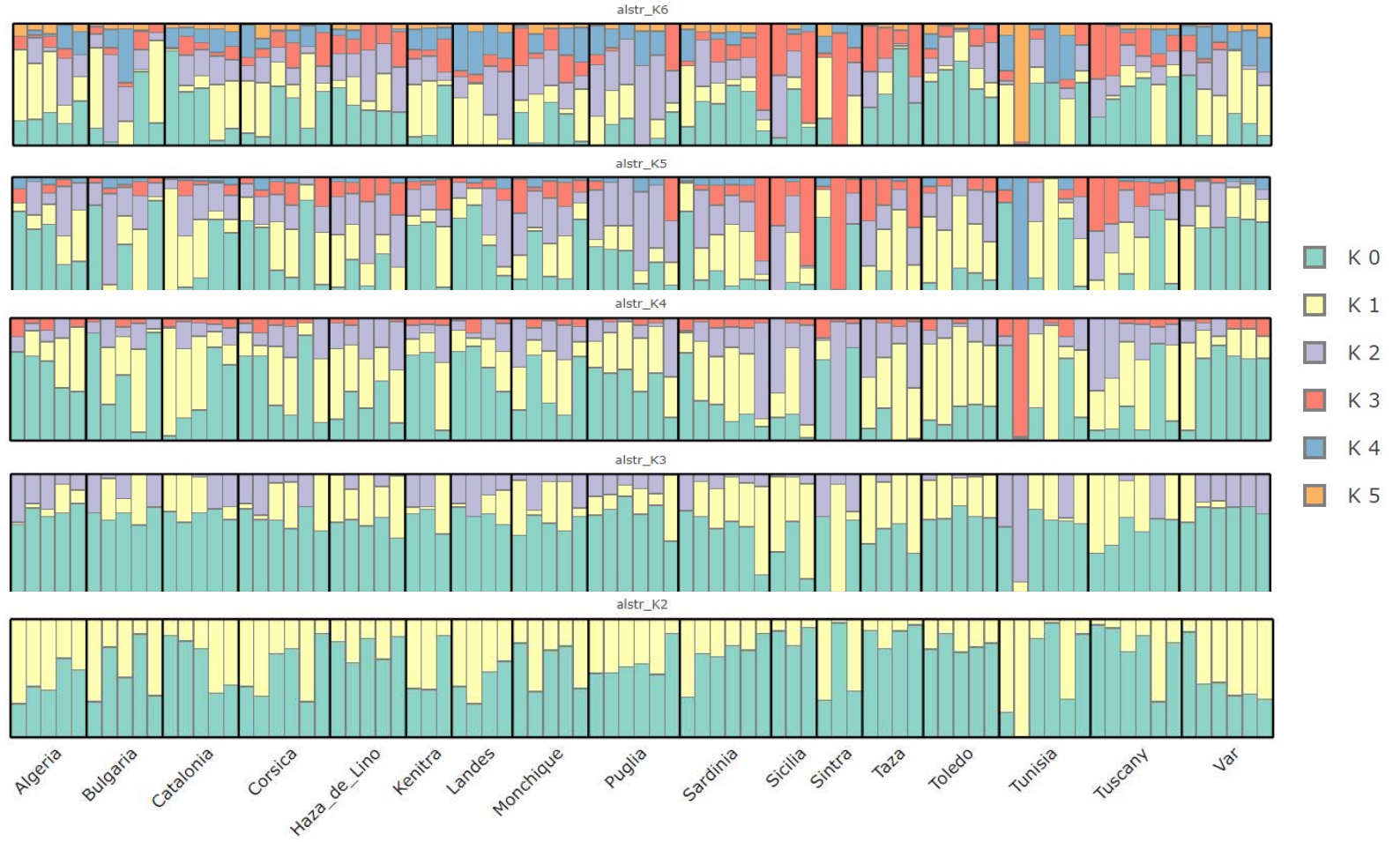

Figura 9 - Gráfico do admixture plot das 83 sequências do sobreiro com K=6 (sendo K=5 o melhor) como está legendado na figura

Com o uso de todos os programas e métodos, referidos anteriormente, foi possível recriarmos uma análise através de um *admixture plot* que tem valores entre K=2 e K=5, ou seja, 4 gráficos diferentes.

Foi importante para a análise, entender que cada barra é um individuo diferente e que cada cluster está identificado com uma cor diferente e que os nossos indivíduos estão agrupados numa amostra com dezassete populações, onde cada uma pertence a uma região diferente à volta do mediterrâneo [FIG.9].

Através de uma análise inicial, podemos observar no gráfico [FIG.9] que, dependendo do valor de K este terá o mesmo número de *clusters* que estão dentro da mistura, ou seja, quando o valor de K é igual a 2, temos na totalidade 2 *clusters*, e se o K for igual a 3 este vai ter 3 *clusters* e por aí a diante.

Iniciando então a nossa análise, a partir de K=2, conseguimos observar que dentro das dezassete localidades estão presentes 2 *clusters*, sendo eles 0 e 1 que apresentam valores de aproximadamente de 50% de cada *cluster*. Já no K=3 com o *cluster* 2 introduzido na amostra observam-se muitas relações entre as nossas espécies devido à presença deste novo *cluster* em todas as amostras em questão.

Considerando o gráfico [FIG.9] com K=4, onde é colocado o *cluster* 3, dentro da mistura, conseguimos identificar uma anomalia no gráfico [FIG.9] relacionado com a amostra da Tunísia, onde uma das barras que representa um indivíduo regista valores próximos de 90% do *cluster* 2, quanto esta foi apresentada no gráfico [FIG.9] de K=3 toma valores diferentes quando o valor de K=4 troca totalmente a barra onde constava o *cluster* 2 passa a valores de *cluster* 3, alterando a cor.

Perante este facto, e verificando que esta anomalia é recorrente durante todo o aumento do valor de K, chegamos à conclusão que, relativamente a um dos indivíduos provenientes da Tunísia, o valor do *cluster* vai sofrendo alterações à medida que é sujeito a diferentes valores de K, ou seja, quando o valor do *cluster* é 2 no K seguinte este passa a tomar valores de *cluster* 3 e assim sucessivamente.

No gráfico com K=5, onde é introduzido o *cluster* 4, a análise fica mais complexa devido à grande aglomeração de dados. Assim, conseguimos concluir que todos os *clusters* são usados em todas as amostras em estudo o que confirma a relação entre as espécies. No entanto, destaca-se que o registo de anomalias referidas anteriormente é constante.

Com base na análise dos diferentes valores de K (K2-K5), onde nos baseamos para decidir o melhor valor de K, decidimos então que o melhor valor de K seria até 5 pois, com a realização dos gráficos com K acima de 5 também é introduzido um novo *cluster* para a nossa amostra fazendo com que a complexidade da mesma fique muito elevada, dificultando a nossa análise. Logo, o valor que consideramos como o mais correto para esta análise foi o gráfico do *admixture plot* com K=5 onde foi possível observar as relações entre estas espécies sendo que todos os valores de *clusters* estão presentes, maioritariamente, em todos os indivíduos de cada espécie, o que justifica estas relações.

Referido anteriormente, foram usadas ferramentas de PCA, ou seja, de redução de dimensionalidade e com elas consegue-se identificar os seus componentes principais. (Lima, Tapadas, & Masson, 2018)

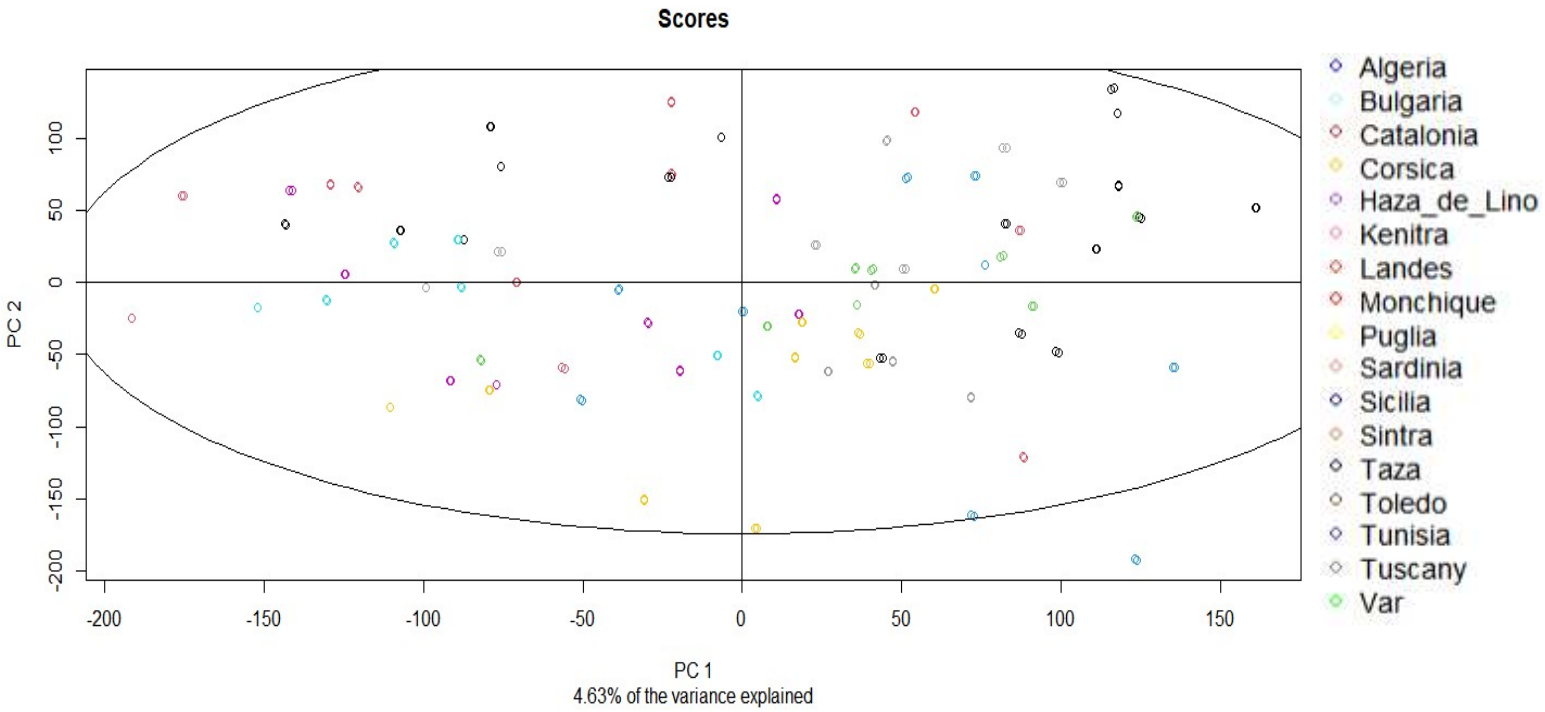

Figura 10 - Gráfico do PCA com os valores de variância expressa do PC1 e PC2 das 83 amostras do sobreiro em que cada circulo (com a sua respetivaa cor) representa uma região como mostrado na legenda do gráfico

Houveram algumas dificuldades para a realização do gráfico [FIG.10] mas com muita pesquisa de vários exemplos dados em sala de aula, as mesmas foram ultrapassadas.

Neste caso, estes fatores são os SNPs de cada amostra. Na Figura 10 – "Gráfico do PCA com os valores de variância expressa do PC1 e PC2 das 83 amostras do sobreiro em que cada circulo (com a sua respetivaa cor) representa uma região como mostrado na legenda do gráfico", o PC1 explica o 4.63% da variação e o PC2 explica o 95.37% da variação.

Cada amostra é representada por um ponto e a região de cada uma delas é representada por uma cor como mostra na legenda. No gráfico do PCA [FIG.10], é possível observar que não há muita junção, ou seja, muita concentração de várias regiões juntas.

Consegue-se ver que a maior parte das regiões estão juntas à respetiva, excepto alguns casos que se consegue observar 2 amostras da *Algeria* fora da zona circulada no gráfico [FIG.10] e outras 2 da mesma região e 2 de *Puglia/Corsica* que se situam na extremidade. Relativamente às amostras de *Taza*, através do gráfico que permite observar que existe uma próximidade com as de Monchique, Bulgária, Var e pode se considerar também *Tuscany* e *Puglia*.

Fora estas observações consegue-se perceber que há mais casos do mesmo tipo ou muito parecidos que são os 2 casos de Tunísia/Haza de Lino e *Corsica*.

Discussão

Com a utilização do *admixture plot* tínhamos como objetivo analisar a estrutura populacional do *Quercus suber* com as populações ancestrais de cada espécie, no sentido de encontrar soluções ou respostas para o nosso problema biológico.

O gráfico do *admixture plot*, quando começa a tomar valores acima de k=3 é possível interpretar que a maioria dos indivíduos apresenta uma relação próxima, uma vez que pertencem todos ao mesmo *cluster*. Esta relação é observada ao longo de toda a análise.

Quanto maior o valor de K, maior será o valor dos *clusters*, fazendo com que a análise do gráfico seja mais complexa e completa. É possível identificar relações de proximidade entre as regiões da amostra como também entre as espécies, contribuindo assim para a nossa interpretação ou conclusão sobre o problema biológico devido a esta proximidade entre espécies.

Comparando com o *admixture plot*, obtido pelos autores (Martins, Batista, Pappas, & Paulo, 2018) do *paper*, tínhamos 3 opções em que a primeira utiliza o conjunto de dados com todos os *loci*, a segunda emprega o conjunto de dados com apenas os *loci* neutros e a outra usa um conjunto de dados com apenas *loci* não neutro.

Perante a opções delineadas, decidimos comparar o nosso *admixture plot* com a primeira opção que utiliza todos os *loci*.

Este *admixture plot* apresentado tem apenas 2 clusters o que nos dá indício que este está relacionado com K=2, que utiliza normalmente 2 *clusters*. Comparado com o nosso gráfico de K=2 é possível observar que os indivíduos estão distribuídos de forma diferente e que relacionado com os *clusters*, estes tomam percentagens diferentes.

Outra diferença está relacionada com a quantidade de indivíduos, pois em alguns casos o número de amostras é diferente: a nossa análise usa oitenta e três amostras enquanto que a dos autores (Martins, Batista, Pappas, & Paulo, 2018) regista um valor de oitenta e um. Assim, concluiu-se que os dois indivíduos removidos são: um da Algeria e outro da Bulgária.

O gráfico do PCA [FIG.10] foi reproduzido apenas pelo nosso grupo de trabalho, pois os gráficos disponibilizados através desta análise que puderam ajudar a retirar mais conclusões e observações para o nosso problema biológico. De tal forma que, não é possivel fazermos uma comparação com o PCA do artigo original.

Contribuições

Alexandre Duarte - Elaboração dos scripts e ajuda no relatório
Diogo Cabrita - Elaboração do relatório e ajuda nos scripts

Referências

Lima, C., Tapadas, N., & Masson, S. (23 de março de 2018). *O que é o PCA – Principal Component Analysis e como aplicá-lo a um conjunto de dados*. Obtido de BI4ALL: https://www.bi4all.pt/noticias/blog/o-que-e-o-pca/

Martins, F., Batista, J., Pappas, G., & Paulo, O. S. (2018). New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Wiley Online Library*, 14.

Schiffel, S., Peltzer, A., & Clayton, S. (2016). *ADMIXTURE analysis*. Obtido de GAWorkshop: https://gaworkshop.readthedocs.io/en/latest/contents/07_admixture/admixture.html

Tiago, P., & Rosário, I. (30 de março de 2019). *Sobreiro* . Obtido de brigadafloresta: https://brigadadafloresta.abae.pt/sobreiro/