IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

Unidade Curricular Análise de Sequências Biológicas

2º ano do curso de licenciatura de Bioinformática

2º Semestre

Ano letivo 2022/2023

Variação genética adaptativa e potenciais estratégias de conservação em Araucaria araucana na região sul da América do Sul

Docente: Francisco Pina

Autores:

Afonso Vaz 202100995

João Venâncio 202100954

Índice

Índice de Figuras

0. Introdução

A *Araucaria araucana*, conhecida como "araucária" ou "*Monkey puzzle tree*", é uma espécie de árvore nativa da região sul da América do Sul. No entanto, como muitas outras espécies, a *Araucaria araucana* enfrenta desafios significativos relacionados às mudanças climáticas e à perda de *habitat*. Compreender a adaptação genética dessa espécie às condições ambientais é crucial para a sua conservação.

O presente trabalho tem como base um estudo científico (Varas-Myrik et al. 2022) que investigou a variação genética adaptativa em *Araucaria araucana* e sua correlação com as condições ambientais. O estudo utilizou uma abordagem de genômica de paisagem, combinando técnicas moleculares e análise estatística para identificar marcadores genéticos adaptativos e avaliar a estrutura populacional da espécie.

O problema biológico original que motivou esta pesquisa é compreender como a *Araucaria araucana* se adapta a diferentes condições ambientais e como essa adaptação pode influenciar sua sobrevivência e persistência diante das mudanças climáticas. Através da genômica de paisagem, os pesquisadores pretenderam identificar os marcadores genéticos adaptativos que conferem uma vantagem adaptativa à espécie em diferentes ambientes.

Para abordar essa questão, os pesquisadores desenvolveram um '*pipeline* bioinformático' específico para o estudo. O '*pipeline* bioinformático' utilizado envolveu a genotipagem das populações de *Araucaria araucana* por meio de técnicas moleculares avançadas. Foi realizada a colheita de amostras de diferentes áreas geográficas e, em seguida, foram realizadas análises genéticas para identificar os marcadores moleculares associados à variação adaptativa.

Os resultados obtidos no estudo revelaram uma alta variação genética adaptativa em *Araucaria araucana*, fortemente relacionada às condições ambientais, como a variação anual da temperatura e os padrões de precipitação. A análise da estrutura populacional revelou a existência de dois grupos genéticos ancestrais com distribuição descontínua entre as cordilheiras dos Andes e da Costa, o que pode ser atribuído a eventos de expansão e retração glacial ao longo do tempo.

Além disso, o estudo identificou fatores limitantes que influenciam a variação genômica adaptativa, como a precipitação do mês mais húmido e do mês mais seco, refletindo compensações fisiológicas associadas a um gradiente de escassez de água. Esses resultados confirmam estudos anteriores que demonstram o papel do *stress* hídrico como um importante fator seletivo na adaptação de árvores. As conclusões do estudo destacam a importância desses resultados para a conservação da *Araucaria araucana*. Sugere-se que as populações localizadas nos Andes sejam priorizadas para a conservação, devido ao alto risco de mal adaptação às

mudanças climáticas futuras. Essas áreas são particularmente sensíveis às mudanças nas condições ambientais e contêm uma grande diversidade genética adaptativa, tornando-as cruciais para a sobrevivência da espécie.

Além disso, o estudo propõe a adoção de estratégias de fluxo genético assistido para auxiliar a *Araucaria araucana* a se adaptar a novos ambientes. O fluxo genético assistido envolve a migração controlada de indivíduos entre populações dentro da faixa da espécie, acelerando o processo de adaptação às condições climáticas futuras. Sugere-se a utilização de uma estratégia de colheita de sementes que misture sementes locais com lotes de sementes pré-adaptadas não locais, visando aumentar a diversidade genética e a resiliência da espécie.

Em resumo, o estudo apresentado neste trabalho aborda a variação genética adaptativa em *Araucaria araucana*, utilizando uma abordagem de genômica de paisagem. Os resultados obtidos fornecem *insights* importantes para a conservação dessa espécie emblemática, destacando a necessidade de considerar a adaptação genética ao lidar com os desafios impostos pelas mudanças climáticas. O *pipeline* bioinformático desenvolvido e os marcadores genéticos adaptativos identificados fornecem uma base sólida para futuras pesquisas e estratégias de conservação eficazes.

1. Objetivos

Neste trabalho, o objetivo principal é investigar a variação genética adaptativa em *Araucaria araucana* e sua relação com as condições ambientais, com o intuito de fornecer informações relevantes para a conservação e controlo dessa espécie emblemática. Para atingir esse objetivo, serão realizadas as seguintes tarefas técnicas:

1. **Coleta e análise dos dados ambientais**: Será realizado um levantamento detalhado das condições ambientais nas áreas de distribuição da *Araucaria araucana*. Serão coletados dados sobre a temperatura, precipitação, altitude e outros fatores relevantes

2. **Realização de análises genéticas**: Serão utilizadas técnicas moleculares avançadas para genotipar as populações de *Araucaria araucana*. Serão identificados marcadores genéticos adaptativos por meio de análises de genômica de paisagem.

3. **Avaliação da estrutura populacional**: Serão realizadas análises para identificar a estrutura populacional da *Araucaria araucana*, que irá envolver a análise da diversidade genética e a investigação de possíveis grupos genéticos ancestrais.

4. **Avaliação dos riscos de mal adaptação e proposição de estratégias de conservação**: Com base nos resultados das análises genéticas e ambientais, serão avaliados os riscos de mal adaptação da *Araucaria araucana* e propostas estratégias de conservação.

2. Materiais e Métodos

Nesta secção, descreveremos os métodos e materiais utilizados para realizar as análises descritas neste trabalho. Explicaremos também os procedimentos que realizámos para cada tarefa incluindo informações sobre o software utilizado, incluindo o número da versão.

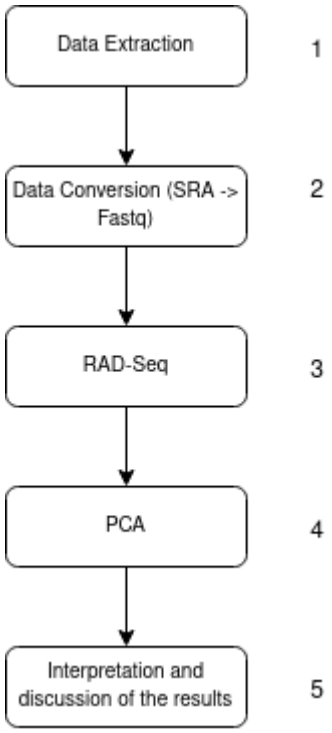


Figura 1 – Pipiline bioinformático

Antes de começarmos, coletámos os dados a partir do Bioproject PRJNA634877, de seguida foi instalado o pacote sra-toolkit (versão 3.0.0) onde o conteúdo do ficheiro tar "sratoolkit.tar.gz" foi extraído. Os dados foram baixados a partir do Bioproject utilizando o comando "prefetch PRJNA634877".

Feito isso, criámos um script "Convert.sh" com o objetivo de converter os arquivos SRA para formato FASTQ. Na preparação dos dados, foi criado um script chamado "Corte.sh" para realizar o corte adicional dos arquivos FASTQ.

2.1. RADSeq

O RADSeq é uma técnica utilizada para criar uma representação reduzida da variação genômica em um conjunto de amostras. Para isso, utilizamos enzimas de restrição para fragmentar aleatoriamente o DNA genômico, seguido por várias etapas possíveis de seleção de fragmentos.

2.1.1 Configuração Inicial

Para iniciar as análises, foi necessário a instalação do Miniconda3 na Virtual Machine (VM). O Miniconda3 é um gerenciador de dependências que nos permitirá instalar o pacote *ipyrad* (v.0.9.92).

2.1.2 Etapas de Montagem

No processo de montagem, a partir da utilização do *software ipyrad* (v.0.9.92) consiste em transformar os dados brutos provenientes do sequenciamento em ficheiro de saída que podem ser utilizados em análises posteriores. As etapas básicas deste processo são as seguintes:

- Etapa 1 - Desmultiplexação e Carregamento dos Dados Brutos
- Etapa 2 - Corte e Controlo de Qualidade
- Etapa 3 - Agrupamento (*clustering*) ou Mapeamento de Referência dentro das Amostras
- Etapa 4 - Cálculo da Taxa de Erro e Heterozigosidade
- Etapa 5 - Chamada de Sequências/alelos de consenso
- Etapa 6 - Agrupamento entre Amostras
- Etapa 7 - Aplicação de Filtros e Geração dos Formatos de Saída

Etapa 1: Desmultiplexação dos dados brutos

Nesta primeira etapa, os sequenciadores fornecem um único arquivo no formato .gz que contém todas as leituras das diferentes amostras misturadas. A etapa de desmultiplexação tem como objetivo separar as leituras pertencentes a cada amostra com base nos códigos de barras. Para isso, utilizamos um ficheiro de códigos de barras que mapeia os nomes das amostras às sequências de códigos de barras correspondentes.

Etapa 2: Filtragem das leituras

Esta etapa envolve a filtragem das leituras com base nas pontuações de qualidade, no número máximo de bases não chamadas e na deteção de adaptadores Illumina. Para lidar com possíveis problemas de contaminação por adaptadores e ruídos nas extremidades das leituras, realizamos um corte nas leituras para um tamanho específico e utilizamos filtros de adaptadores mais agressivos.

Etapa 3: Agrupamento dentro das amostras

Nesta etapa, as leituras são agrupadas *(clustering*) dentro de cada amostra, considerando um início de similaridade definido. O objetivo é identificar as leituras que mapeiam para o mesmo loco dentro de cada amostra.

Etapa 4: Estimativa conjunta de heterozigosidade e taxa de erro

Esta etapa envolve a estimativa conjunta da taxa de heterozigosidade e da taxa de erro. Detalhes adicionais sobre os métodos específicos utilizados para essa estimativa podem ser encontrados na documentação do *ipyrad* (v.0.9.92).

Etapa 5: Chamada de sequências/alelos de consenso

Nesta etapa, são realizadas várias operações, incluindo o cálculo das profundidades, o agrupamento em pedaços menores para melhorar a paralelização, a chamada de sequências de consenso e o indexamento dos alelos.

Etapa 6: Agrupamento entre as amostras

Nesta etapa, as sequências de consenso são agrupadas entre as amostras com base em um limiar de similaridade. Isso envolve a concatenação dos arquivos de entrada, o agrupamento por similaridade, a construção dos agrupamentos e o alinhamento das sequências dentro de cada agrupamento.

Etapa 7: Filtragem e geração de arquivos de saída

Por último, na etapa final consiste na filtragem dos dados e na geração dos ficheiros de saída nos formatos desejados. No caso deste trabalho, os formatos de saída especificados no ficheiro de parâmetros resultarão em arquivos nos formatos phylip, VCF e loci.

2.2 Structure_threader

Feito o RADSeq com uso do *software ipyrad* das sete etapas que foram anteriormente mencionadas, vamos recorrer ao *software* para realizar análises de estrutura populacional foi o '*Structure_threader'*, que envolve três programas diferentes para realizar análises de estrutura em paralelo: *STRUCTURE*, *fastStructure* e *ALStructure*, fornecendo uma interface unificada para executar esses programas e realiza várias execuções em paralelo.

Após a instalação do '*Structure_threader'*, foram realizadas as etapas necessárias para obter os dados e prepará-los para a análise de estrutura populacional. Foi baixado um ficheiro 'VCF' contendo os dados genéticos e foi criado à mão um ficheiro 'indfile' contendo informações sobre os indivíduos e as localizações. Este ficheiro 'indfile' é um ficheiro de texto no qual a primeira coluna representa o nome da amostra individual e a segunda coluna representa o código de localização onde ela foi coletada, fornecendo informações essenciais para associar as amostras genéticas aos respetivos locais geográficos de origem. Vai ser utilizado para rastrear a

proveniência dos indivíduos e auxiliar na análise de padrões de diversidade genética em diferentes populações ou regiões., sendo essas informações recolhidas de uma tabela (Figura 1) do paper em estudo (Varas-Myrik et al. 2022).

Map code (Fig. 1)	Site name	Latitude	Longitude	N (Valid)	Mountain range
BP	Bosque Pehuén	−39.46	−71.72	12	Andes
HL	Hualalafquén	−39.33	−71.41	7	Andes
LM	La Mula	−37.90	−71.37	11	Andes
LR	Las Raíces	−38.43	−71.45	14	Andes
MC	Malalcahuelo	−38.43	−71.54	8	Andes
MM	Mamuil Malal	−39.58	−71.48	10	Andes
PC	Conguillío	−38.70	−71.81	9	Andes
PN	Parque Nahuelbuta	−37.81	−73.02	14	Costa
RA	Ralco	−37.94	−71.34	14	Andes
TH	Tolhuaca	−38.20	−71.78	9	Andes
TR	Trongol	−37.69	−73.12	11	Costa
VA	Villa Araucarias	−38.49	−73.25	15	Costa

Figura 2- Localizações de amostragem, coordenadas geográficas (graus de latitude sul e longitude oeste), tamanho da amostra (N) e Cordilheira das 12 áreas de amostragem de A. araucana.

Além disso, foram realizadas etapas de filtragem de frequência alélica mínima e desequilíbrio de ligação.

Embora não termos realizados os seguintes passos, iriam prosseguir da seguinte forma:

- Em seguida, o '*Structure_threader'* seria executado com o programa *ALStructure* usando o comando *structure_threader run*. Especificaria-se parâmetros como o arquivo de entrada (-i), o diretório de saída (-o), o número de clusters a serem testados (-K), o número de núcleos de CPU a serem usados (-t) e o arquivo de informações individuais (--ind).
- Após a conclusão da execução do *Structure_threader*, seriam gerados arquivos de saída, incluindo gráficos de *admixture*, sendo esses guardados em formatos vetoriais (SVG) e também num formato interativo (HTML) que nos permitiria explorar os resultados de forma dinâmica.

2.3 Introdução à Análise de Componentes Principais (PCA)

De seguida, recorremos ao PCA utilizando o software *Rstudio (v.4.1.2)* onde foram utilizados os seguintes *packages: "LEA", "pcaMethods", "vcfR" e "ggplot2"*. No âmbito da criação do PCA, foi reutilizado o 'indfile' criado anteriormente.

O PCA é uma técnica estatística utilizada para reduzir a dimensionalidade dos conjuntos de dados multivariados, permitindo uma representação mais compacta e interpretação dos padrões subjacentes nos dados, permitindo-nos assim ver onde os indivíduos são posicionados no espaço bidimensional de acordo com suas características genéticas e permitindo também identificar padrões de similaridade ou diferenciação entre as populações de diferentes localizações geográficas.

3. Resultados

3.1 Estrutura populacional

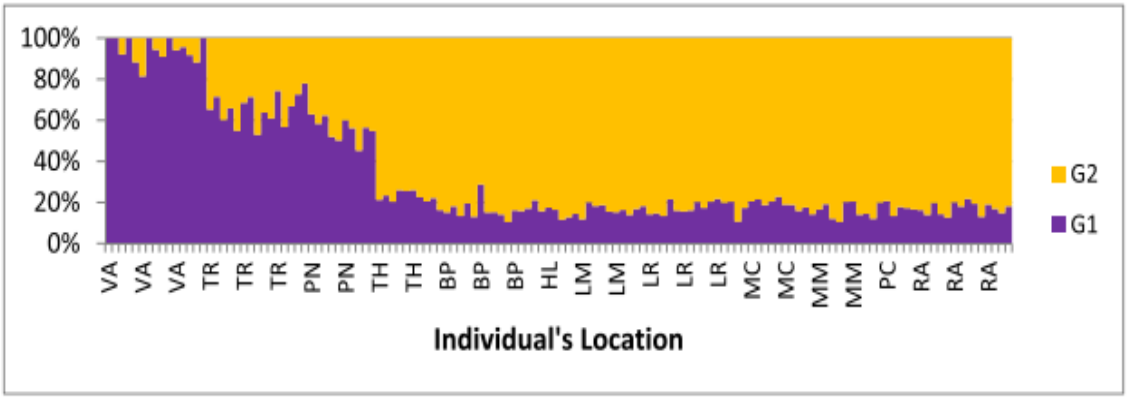


Figura 3 - Afiliações populacionais (K = 2) dos 134 indivíduos de A. Araucana, com base no algoritmo sNMF. As barras verticais representam a percentagem do genoma de cada indivíduo que pertence a cada um dos dois grupos ancestrais (G1 e G2). Observa-se que os indivíduos da localização VA representam uma classe distintiva de mistura, com uma presença elevada de G1. Os indivíduos de TR e PN constituem uma segunda classe de mistura, com uma presença relativamente equilibrada de G1 e G2, enquanto os restantes indivíduos da Cordilheira dos Andes formam uma terceira classe de mistura, com uma baixa presença de G1.

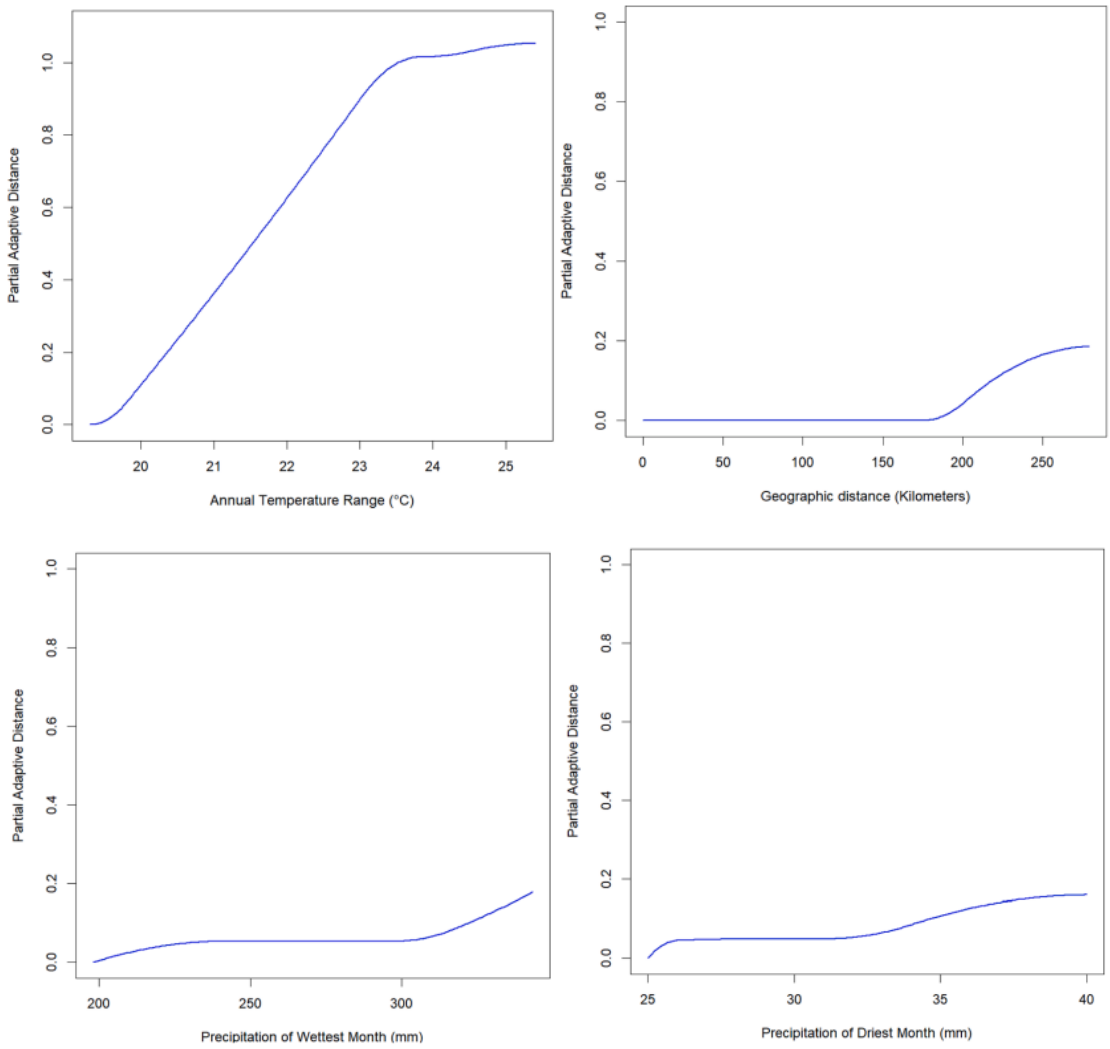
Embora não termos realizado *o admixture plot*, podemos sempre fazer uma análise deste criado no paper em estudo (Varas-Myrik et al. 2022).

A aplicação do algoritmo sNMF permitiu identificar a estrutura populacional da espécie *Araucaria araucana*. A análise revelou a presença de dois grupos ancestrais distintos, representados pelos clusters G1 e G2. A distribuição dos coeficientes de ancestralidade

demonstrou uma clara diferenciação entre as populações localizadas nas Cordilheiras dos Andes e aquelas localizadas na região costeira (Fig. 2).

No caso das populações Andinas, foi observada uma homogeneidade nos níveis de mistura, com aproximadamente 20% dos indivíduos pertencentes ao *cluster* G1 e 80% ao *cluster* G2. Por outro lado, as populações costeiras apresentaram níveis heterogêneos de mistura ancestral. Os indivíduos das localidades PN e TR contribuíram com cerca de 60% do *cluster* G1 e 40% do *cluster* G2, enquanto os indivíduos de VA apresentaram níveis de mistura em torno de 90% do *cluster* G1 e 10% do *cluster* G2.

3.2 Adaptive divergence model (GDM) fit



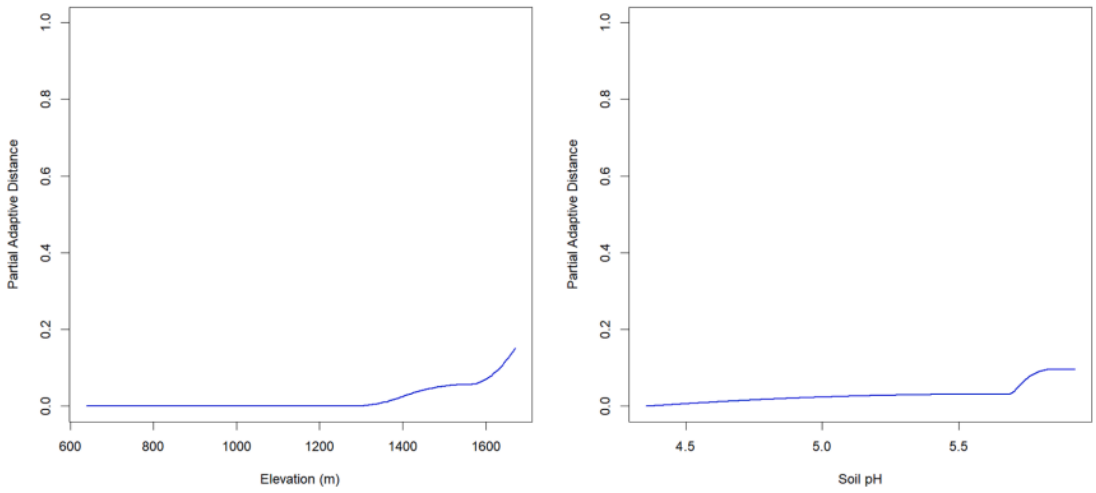


Figura 4 - Gráficos GDM I-splines para cada variável ambiental e distância geográfica. A altura máxima de cada curva indica a quantidade total de alteração nas frequências alélicas associadas a essa variável (importância da variável). A forma de cada curva indica como a taxa de alteração nas frequências alélicas varia ao longo do gradiente da variável.

Embora não termos realizado os Gráficos GDM, podemos sempre fazer uma análise destes criados no paper em estudo (Varas-Myrik et al. 2022).

O modelo GDM explicou 85,1% da variação na distância genética adaptativa individual. A altura máxima de cada I-Spline (Fig. 3) indica a quantidade total de alteração nas frequências alélicas associadas a essa variável (importância da variável). A faixa anual de temperatura mostrou ser a variável mais importante, seguida pela distância geográfica, precipitação do mês mais húmido, precipitação do mês mais seco, elevação e pH do solo. Em relação à forma das I-Splines (Fig. 3), a faixa anual de temperatura apresentou um padrão linear entre 19 e 23 °C, após o qual mostrou uma diminuição acentuada na inclinação, sugerindo um comportamento de estabilização após esse limiar. A distância geográfica mostrou um padrão diferente: entre 0 e 160 km, a sua importância foi relatada como nula, mas acima de 160 km, a importância aumentou constantemente

3.3 Introdução à Análise de Componentes Principais (PCA)

No gráfico de PCA (Figura 5), cada ponto no gráfico representa um indivíduo e a cor do ponto indica a localidade em que o indivíduo foi colheitado. A análise de PCA nos permitiu visualizar a distribuição dos indivíduos ao longo dos componentes principais, identificar agrupamentos ou separações entre as localidades e investigar a existência de estrutura populacional.

Figura 5 – Representação gráfica da análise de Componentes Principais (PCA) para a população de Araucaria araucana em diferentes localizações.

Contudo, como no paper (Varas-Myrik et al. 2022) não apresenta uma representação gráfica de PCA, não será possível fazer uma a comparação entre os nossos resultados com os deles.

4. Discussão

4.1 – PCA

Ao observar o PCA (Figura 5), podemos identificar padrões de similaridade ou diferenciação entre os indivíduos com base na sua proximidade espacial no espaço bidimensional definido pelos dois componentes principais (PC1 e PC2) calculados na análise de PCA.

Caso os pontos de indivíduos de uma mesma localidade estejam agrupados numa área específica do gráfico, pode-se então sugerir que esses indivíduos compartilham características genéticas semelhantes. Por outro lado, se os pontos de indivíduos de diferentes localidades se encontram espalhados por áreas distintas do gráfico pode-se verificar que esses indivíduos têm diferenças genéticas significativas. Além disso, a distância entre os pontos no gráfico de PCA também é relevante, pois os pontos que estão próximos uns dos outros indicam maior similaridade genética, enquanto pontos distantes estão mais geneticamente distintos.

Ao identificar populações ou grupos de indivíduos com características genéticas distintas em diferentes localidades, podemos inferir que essas populações possuem adaptações específicas às condições ambientais locais. Isso é crucial para entender como a espécie responde às pressões seletivas em diferentes regiões e quais são os fatores ambientais que desempenham um papel significativo na sua adaptação.

Essas informações podem ser aplicadas no desenvolvimento de estratégias de conservação e manejo da espécie. Ao reconhecer a existência de populações geneticamente distintas, podemos adotar abordagens de conservação mais direcionadas, levando em consideração as particularidades genéticas e adaptativas de cada população. Isso pode incluir a implementação de medidas de conservação específicas para cada região, considerando as necessidades e as ameaças específicas enfrentadas por cada população.

Além disso, a compreensão da estrutura populacional da espécie por meio do PCA pode auxiliar na identificação de áreas prioritárias para a conservação. Ao identificar populações geneticamente distintas e áreas de alta diversidade genética, podemos direcionar os esforços de conservação para regiões que desempenham um papel fundamental na manutenção da variabilidade genética da espécie.

Essas conclusões fornecem-nos insights valiosos sobre a adaptação geográfica da espécie em estudo. Essas informações contribuem para uma melhor compreensão da diversidade genética e da estrutura populacional da espécie, bem como para futuras investigações sobre a evolução e conservação dessas populações.

4.2 – Avaliação dos riscos e propostas de estratégias de conservação

Com base nos resultados das análises genéticas e ambientais, que incluem informações sobre a estrutura genética da população, os padrões de adaptação geográfica e a influência dos fatores ambientais na diversidade genética, é possível identificar potenciais riscos de mal adaptação da espécie.

Esses riscos podem ser decorrentes de ameaças como a perda de habitat, a fragmentação do ambiente, as mudanças climáticas e a introdução de espécies invasoras, escassez de precipitação, entre outros fatores. A análise genética e a compreensão da adaptação geográfica da espécie fornecem *insights* valiosos para avaliar a capacidade de resposta da *Araucaria araucana* a essas ameaças.

Com base nessa avaliação de riscos, podem ser propostas estratégias de conservação adequadas para mitigar os impactos negativos e promover a adaptação da espécie. Essas estratégias podem incluir a proteção e recuperação de habitats-chave, a implementação de medidas de controlo para reduzir a fragmentação do ambiente, a promoção da conectividade entre populações e ações para minimizar os efeitos das mudanças climáticas.

5. Referências

BioSample for BioProject (Select 634877) - BioSample - NCBI. (s.d.). Obtido de National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/biosample?Db=biosample&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_biosample&LinkReadableName=BioSample&ordinalpos=1&IdsFromResult=634877

Pina, F. (s.d.). *Usage - Structure_threader manual*. Obtido de Introduction - Structure_threader manual: https://structure-threader.readthedocs.io/en/latest/usage/#using-a-popfile

saketkc. (s.d.). *GitHub - saketkc/pysradb: Package for fetching metadata and downloading data from SRA/ENA/GEO*. Obtido de GitHub: https://github.com/saketkc/pysradb

Varas-Myrik, A. (15 de 01 de 2022). *Predicting climate change-related genetic offset for the endangered southern South American conifer Araucaria araucana* . Obtido de ScienceDirect: https://www.sciencedirect.com/science/article/pii/S0378112721009476