

DIVERSIDADE GENÔMICA E ESTRUTURA GENÉTICA DE PEIXES TELEÓSTEOS PARA O DESENVOLVIMENTO DE ESTRATÉGIAS DE CONSERVAÇÃO DE RECURSOS PESQUEIROS ASSOCIADOS A PESCA ARTESANAL¹

Isis Ferreira Silveira de Souza², Miklos Maximiliano Bajay³

¹ Vinculado ao projeto “Diversidade genômica e estrutura genética de peixes teleósteos para o desenvolvimento de estratégias de conservação de recursos pesqueiros associados a pesca artesanal”

² Acadêmico (a) do Curso de Ciências Biológicas - CERES – Bolsista PROBIC

³ Orientador, Departamento de Engenharia de Pesca e Ciências Biológicas – CERES – miklos.bajay@udesc.br

O desenvolvimento tecnológico permitiu o sequenciamento dos genomas de diversos organismos e vem progredindo de forma exponencial, permitindo a exploração de uma infinidade de possibilidades de conservação, manejo e prevenção de doenças e melhoramento genético. A tecnologia avançou tanto que, hoje, o sequenciamento é apenas a base dos estudos e gera desafios de interpretação relacionados à geração de uma quantidade massiva de dados. Através da bioinformática podemos usufruir integralmente da tecnologia disponível, tendo em vista que os dados gerados no sequenciamento precisam ser manipulados e analisados através de ferramentas computacionais.

VCF é um formato padrão dos dados de sequenciamento usado em bioinformática e consiste em um sistema simplificado para armazenar e visualizar informações de variação genética. O presente estudo tem como objetivo apresentar e discutir algumas das principais ferramentas de análise genômica com base em arquivos VCF.

Os arquivos VCF podem ser manipulados e analisados no R através do pacote adegenet, dedicado a análises multivariadas de dados de marcadores genéticos, que conta com gráficos, testes estatísticos, ferramentas de simulação, medidas de distância e similaridade e vários métodos espaciais, além de fornecer vários conjuntos de dados simulados e empíricos.

O arquivo VCF pode ser transformado em um vetor através do pacote vcfR e esse vetor pode ser transformado nas classes de objetos S4 do pacote adegenet, para que possa ser analisado. Objetos são unidades que interagem entre si na Programação Orientada a Objetos (POO), podendo ser divididos em classes que determinam comportamento e estados possíveis de seus objetos, além da forma como se relacionam com outros objetos. O padrão de orientação a objeto é originado nos estudos da cognição e tem forte apelo na redução do gap semântico entre o mundo real e a representação dele, de forma que os conceitos sejam traduzidos em ideias menos abstratas e mais intuitivas, já que objetos têm características e ações, assim como vemos no mundo real.

O R é uma linguagem de programação destinada à visualização, manipulação e análise de dados, que possui três sistemas orientados a objetos (OO): S3, S4 e R5. Muitas das ideias básicas de objetos S3 são comuns a S4, mas em S4 a implementação é muito mais estrita. Os objetos do pacote adegenet, genind, genpop e genlight, são de classe S4 e armazenam, respectivamente, informações de genótipos individuais, de genótipos de grupos de indivíduos e de dados de SNPs para todo o genoma.

O objeto de classe S4 `genind` é utilizado para armazenar genótipos individuais através da criação de um objeto `genind` a partir de uma tabela de alelos, a qual corresponde ao slot `@tab`. O slot `@tab` trata-se de valores inteiros dispostos em uma matriz que contém dados de genótipos para indivíduos (linhas) para todos os alelos (colunas). O slot `@type` define o tipo de marcador: “codom” para marcadores codominantes, nos quais os valores são números de alelos e “PA”, nos quais os valores são a presença ou ausência de alelos.

Os arquivos VCF podem ser manipulados no R através do pacote `vcfR`, que possibilita criar vetores a partir da leitura do arquivo de interesse, utilizando a função `read.vcfR()`.

Para transformar o vetor “vcf” em `genind`, basta usar a função `vcfR2genind()`. Em seguida pode-se remover o vetor `vcf` a fim de otimizar o uso da memória RAM na continuidade da análise.

Tabela 1. Colunas obrigatórias que descrevem os dados VCF.

Sigla	Significado	Descrição
CHROM	Chromosome	Identificador da sequência na qual se encontra a variação
POS	Position	Posição da variação na sequência, primeira base representa a posição 1
ID	Identifier	Lista separada por ponto e vírgula de identificadores únicos, quando disponíveis. Nenhum identificador deve estar presente em mais de um registro de dados. Se não houver identificador disponível, (“.”) deve ser usado
REF	Reference base(s)	A, T, C, G ou N (independente de maiúsculas e minúsculas). Bases múltiplas são permitidas.
ALT	Alternate base(s)	Lista dos alelos alternativos à referência, separados por vírgula. A representação “*” indica ausência do alelo devido a deleção
QUAL	Quality	Pontuação de qualidade de acordo com a escala Phred para a afirmação feita em ALT
FILTER	Filter status	PASS se uma chamada é feita nesta posição e ela passou todos os filtros. Se o local não passou todos os filtros, há uma lista de códigos para os filtros que falharam, separados por ponto e vírgula
INFO	Additional information	Campos INFO são codificados como uma série de short Keys separados por ponto e vírgula. Exemplo: NS=59;AF=0.102 (onde NS é o número de amostras com dados e AF é a frequência alélica para cada alelo ALT)

```

4 library(vcfR)
5 vcf <- read.vcfR(file = "robalo.vcf", verbose = FALSE)
6 vcf <- read.vcfR(file = "plink.vcf", verbose = FALSE)
7 input.genind <- vcfR2genind(vcf, return.alleles = TRUE)
8 remove(vcf)
9 indNames(input.genind)
10 indNames
11 summary(input.genind)
12 library(adegenet)
13 library(poppr)
14 pop(input.genind) <- c(rep("undecimalis", 24), rep("parallelus", 7), rep("undecimalis", 7))
15 inputNA.genind <- missingno(input.genind, type = "mean", cutoff = 0.00, quiet = TRUE)
16 pob <- data.frame(nom=indNames(inputNA.genind), pob=pop(inputNA.genind))

```

Figura 1. Script executado na pesquisa

Palavras-chave: VCF. R. adegenet.