

UNIVERSIDADE ESTADUAL PAULISTA (UNESP)
FACULDADE DE CIÊNCIAS AGRÁRIAS E TECNOLÓGICAS
CAMPUS DE DRACENA

RENAN FABRÍCIO GONELA

DESENVOLVIMENTO DE FUNÇÕES DE SUPORTE À ANÁLISES
GENÉTICAS/GENÔMICAS PARA O SISTEMA R

Dracena

2022

UNIVERSIDADE ESTADUAL PAULISTA (UNESP)
FACULDADE DE CIÊNCIAS AGRÁRIAS E TECNOLÓGICAS
CAMPUS DE DRACENA

Renan Fabrício Gonela

DESENVOLVIMENTO DE FUNÇÕES DE SUPORTE À ANÁLISES
GENÉTICAS/GENÔMICAS PARA O SISTEMA R

Trabalho de Conclusão de Curso apresentado à
Faculdade de Ciências Agrárias e Tecnológicas
– Unesp, Câmpus de Dracena como parte das
exigências para conclusão do curso em
Zootecnia.

Orientador: Prof. Associado Ricardo da Fonseca
Coorientador: Alexandre de Queiroz

Dracena
2022



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Dracena



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRÁRIAS E TECNOLÓGICAS
UNESP – CÂMPUS DE DRACENA

CERTIFICADO DE APROVAÇÃO

Título: Desenvolvimento de funções de suporte à análises genéticas/genômicas para o sistema R

Modalidade: Trabalho de Atividades de pesquisa

Autor:

Orientador: Prof. Associado Ricardo da Fonseca

Co-orientador: Alexandre de Queiroz

Número de Créditos: 15

Data da aprovação e correção de acordo com as sugestões da Banca: 15 / 12 / 2022

Prof. Dr. Celso Tadao
Miasaki
Nome membro da Banca

Prof. Dr. Gustavo do Valle
Polycarpo
Nome membro da Banca

Prof. Dr. Ricardo da
Fonseca
Nome membro da Banca

DADOS CURRICULARES DO AUTOR

Renan Fabrício Gonela, nascido em 26 de Abril de 1993, na cidade de Tupã/SP. Filho do pai Joel Gonela e da mãe Sônia Maria da Graça Gonela. Discente do curso de Zootecnia, professor bolsista do Cursinho Pré-Vestibular Alvo desde o 1º Semestre de 2020, ministrando aulas na disciplina de Química Orgânica, atual monitor da disciplina de Métodos de Melhoramento Animal, ex-monitor da disciplina de Microbiologia Zootécnica, e ex-bolsista do programa Bercas Santander Graduações – 2º Semestre 2019 - Excelência acadêmica.

DEDICATÓRIA

Ao meu pai **Joel Gonela** e minha mãe **Sonia Maria da Graça Gonela**, que empenharam muito esforço e apoio para que eu pudesse concluir minha graduação com rendimento satisfatório, esta conquista é dedicada a eles.

AGRADECIMENTOS

Dedico este trabalho de conclusão de curso para:

Meu pai, Joel, que desde o momento que deixei meu trabalho estável para ir em busca do meu sonho em ser Zootecnista me apoiou, dando amparo nas minhas decisões, nos momentos que voltei para a casa sempre conversávamos sobre o nosso time de coração e como as coisas estavam, aos momentos do churrasco dos domingos, depois de todas as viagens cansativas, me recebia de braços abertos com a minha mãe.

Minha mãe, um exemplo de mulher e mãe, a “dona Sônia”, que sempre dedicou tempo na minha educação e na educação do meu irmão, que me atentava aos riscos dos caminhos fora da escola, que por vezes puxava a minha orelha, hoje entendo todo o seu esforço, sou muito grato por tudo.

Ao meu irmão Rafael, um dos meus maiores incentivadores, um exemplo, agradeço pelos conselhos durante a minha vida, pela atenção, e que recentemente me apresentou com o Vicente, meu afilhado, que mudou a minha vida.

À minha família, que sempre me acolhia nos momentos de lazer e relaxamento.

À Juliana, que esteve comigo nos momentos bons e ruins, me apoiando e aconselhando.

Aos meus amigos, pelas conversas e conselhos, devo citar o Gabriel e o Eduardo, mesmo que distantes sempre tiveram atenção na nossa amizade, ao Leandro o “Lê” e ao meu primo Luan, que mesmo partindo recentemente sempre serão carregados no meu coração.

Aos meus animais de estimação, que fazem parte da minha família, que de maneira única e especial me confortavam com todo carinho.

Ao professor Dr. Ricardo da Fonseca, meu orientador, pela atenção e paciência, pelo acolhimento e conhecimentos na área do Melhoramento Genético Animal.

Ao Cursinho Pré-Vestibular Alvo, que nestes anos me fez enxergar a importância de um projeto de extensão, que impactou a minha vida e me fez crescer como pessoa e futuro profissional, aos alunos que ministrei aulas de Química Orgânica, guardo imenso carinho.

A todos os professores e colaboradores da Faculdade de Ciências Agrárias e Tecnológicas – Unesp, Câmpus de Dracena, pela atenção durante a minha formação.

Aos membros do grupo de estudo LuCCA-Z, onde dividimos momentos incríveis.

Ao meu grande mestre Fábio, que durante o ensino fundamental e médio me orientava sobre a importância do ensino público.

A todos que passaram pela minha vida acadêmica, muito obrigado.

“É necessário sempre acreditar que o sonho é possível...” (Racionais Mc's, 2002).

RESUMO

Nos programas de Melhoramento Genético Animal, estimação de parâmetros genéticos e valores genéticos aditivos são rotina e um passo muito importante para o processo de seleção e construção de estratégias de acasalamentos. Entretanto, para que os ganhos genéticos sejam significativos, essas estimativas precisam ser feitas com qualidade. Nesse sentido, é fundamental trabalhar a consistência dos dados, a formação consciente de novas variáveis como grupos de contemporâneos e a formatação correta do arquivo de dados para entrada em programas de avaliação genética. Essas análises preliminares, de modo geral, tomam muito tempo e não seguem um padrão básico. O uso de pacotes computacionais facilitam, poupam tempo e aumentam a eficiência das análises, minimizando erros durante o processo. Assim, nosso objetivo foi dar início a criação de um pacote para ser utilizado no software R que possuem funções que automatizam a avaliação dos dados, e geram informações estratégicas ao usuário. Neste trabalho foi utilizado o software R para a elaboração de funções para consistência de dados, formação de grupo de contemporâneos e formatação dos dados para o software de avaliações genéticas Wombat. Todas as funções foram colocadas dentro de um pacote denominado LZ13.

Palavras-chave: Melhoramento genético animal. Software R. Consistência de dados. Formatação de dados. Funções para o R. Pacotes para o R.

ABSTRACT

In Animal Genetic Improvement programs, patterns of genetic patterns and additive genetic values are routine and a very important step in the process of selection and construction of mating strategies. However, for the genetic gains to be experienced, these estimates need to be done with quality. In this sense, working on data consistency, the conscious formation of new variables such as contemporary groups and the correct formatting of the data file for input into genetic evaluation programs, is fundamental. These preliminary analyzes generally take a long time and do not follow a basic pattern. The use of computational packages facilitates, saves time and increases the efficiency of the analysis, minimizing errors during the process. Thus, our goal was to start creating a package to be used in the R software that has functions that automate data evaluation and generate strategic information for the user. In this work, the R software was used to create functions for data consistency, formation of contemporaneous groups and data formatting for the Wombat genetic estimation software. All functions were placed inside a package called LZ13.

Keywords: Animal genetic improvement. R software. Data consistency. Data formatting. Functions for R. Packages for R.

LISTA DE FIGURAS

Figura 1 – Função moda.....	24
Figura 2 - Exemplo da moda.....	24
Figura 3 – Função de consistência dos dados.....	25
Figura 4 – Exemplo da função consistF.....	26
Figura 5 – Gráficos de frequências da função consistF para variável idamae_par..	26
Figura 6 – Função para grupo de contemporâneos.....	27
Figura 7 – Exemplo da formação de grupos de contemporâneos.....	28
Figura 8 – Função auxiliar de formatação dos dados para o Wombat.....	29
Figura 9 – Exemplo da função Wombat.....	30
Figura 10 - Descrição do pacote LZ13.....	31
Figura 11 - Funções do pacote LZ13.....	31

LISTA DE TABELAS

Quadro 1 - Comparação entre pacotes existentes para R na área de melhoramento genético animal e o pacote proposto no presente trabalho	20
---	----

LISTAS DE ABREVIATURAS E SIGLAS

AGHmatrix	Relationship Matrices for Diploid and Autopolyploid Species
gggroup	Pedigree and Genetic Groups
GHap	Genome-Wide Haplotyping
hiphop	Parentage Assignment using Bi-Allelic Genetic Markers
MDSMap	High Density Genetic Linkage Mapping using Multidimensional Scaling
optiSel	Optimum Contribution Selection and Population Genetics
phangorn	Phylogenetic Reconstruction and Analysis
phyclust	Phylogenetic Clustering
purgeR	Inbreeding-Purging Estimation in Pedigreed Populations
QGglmm	Estimate Quantitative Genetics Parameters from Generalised Linear Mixed Models
SFSI	Sparse Family and Selection Index
sommer	Solving Mixed Model Equations in R
STGS	Genomic Selection using Single Trait
xbreed	Genomic Simulation of Purebred and Crossbred Populations
DEP	Diferença esperada na progênie
TPA	Predicted transmitting ability
CRAN	The Comprehensive R Archive Network

SUMÁRIO

1 INTRODUÇÃO	15
2 OBJETIVOS	16
2.1 Objetivos Específicos	16
3 REVISÃO DE LITERATURA	17
3.1 Software R e seus pacotes.....	17
3.2 Ferramentas para desenvolvimento de Pacotes em R.....	18
3.3 Pacotes para Melhoramento Genético Animal em R.....	18
4 MATERIAL E MÉTODOS	20
5 DESENVOLVIMENTO	21
6 CONCLUSÃO.....	31
7 REFERÊNCIAS.....	32

1 INTRODUÇÃO

A área de Melhoramento Genético Animal teve grande avanço com o aparecimento dos computadores, pois, até então, pela necessidade de manipulação de grandes quantidades de dados e cálculos complexos, tinha sua aplicação e contribuição à produção animal limitada.

A formulação de estratégias de ação em programas de melhoramento genético e ações de pesquisa se baseiam na análise de grandes quantidades de dados. Mais recentemente, com a coleta automatizada de fenótipos e dados genômicos, a área de melhoramento tem trabalhado sob o contexto Big Data, onde enormes volumes de dados devem ser manipulados.

Todos os procedimentos necessários para o trabalho com os dados só podem ser executados com o uso de computadores e softwares especializados.

Toda análise de dados demanda, como primeiro passo, independente da natureza dos registros, uma análise de consistência, a qual tem como objetivo aumentar a qualidade dos dados e, conseqüentemente, proporcionar resultados confiáveis e consistentes ao final das análises.

No melhoramento genético animal, a consistência dos dados deve ser feita para várias categorias de dados: fenotípicos, pedigree e genômicos. Para cada categoria, o racional da análise de consistência muda completamente, dando origem a análises completamente diferentes.

Uma vez que os dados tenham passado pela etapa de consistência, esses devem seguir para os softwares de análise genéticas, onde vários aplicativos estão disponíveis para as possíveis análises.

Cada software demanda uma formatação específica dos dados a serem utilizados como entrada, demandando muito cuidado e muito trabalho por parte do utilizador/pesquisador. Na prática, a etapa de ajuste e formatação dos dados para entrada nos softwares é uma etapa que consome considerável tempo de trabalho.

Atualmente o software R tem sido muito utilizado pelos profissionais da área de melhoramento genético animal para a manipulação, formatação e análise de dados.

Entretanto, assim como ocorre nos simuladores de dados, as soluções desenvolvidas por meio de scripts e funções são particulares e não disponibilizadas publicamente e de forma organizada para a utilização de terceiros, principalmente se pensarmos em soluções e documentação na língua portuguesa.

A organização e disponibilização das soluções para problemas gerais como a consistência de dados fenotípicos, formação de grupos de contemporâneos, consistência e recodificação do pedigree, controle de qualidade de dados genômicos e formatação de arquivos para entrada em análises genéticas, seria muito

interessante para o ensino e desenvolvimento de pesquisas com o software R, principalmente ao nível de graduação e mestrado, que tem como característica, curtos intervalos de tempo para a realização do TCC e dissertação, respectivamente.

Nesse sentido, os alunos não precisariam se preocupar com construção de cada fragmento de código que seriam necessários em suas análises, uma vez que todo o código já estaria contido em uma função, a qual seria disponibilizada pela instalação do pacote no sistema base.

Ao ativar a função grande parte do trabalho já seria feito, e permite ao aluno focar mais nos resultados do que no processo de desenvolvimento.

A mudança de foco agiliza o trabalho, sendo bastante benéfica aos alunos de graduação e mestrado, que, de modo geral, têm pouco tempo para preparar seus trabalhos. Além disso, a utilização de funções torna o trabalho mais leve, evitando que os alunos precisem se aprofundar na linguagem R para a execução de suas tarefas.

Sob o ponto de vista do orientador, muito trabalho é evitado ao se utilizar as funções que estão organizadas e já foram testadas com novos alunos. Não é necessário que explicar todo o processo de desenvolvimento novamente e para os alunos, o entendimento dos processos fica facilitado.

Assim, as funções possibilitam que o aprendizado do software se dê de maneira gradual, com a utilização de elementos mais básicos no curto período disponível.

A organização de diversas rotinas para consistência e formatação dos dados pode auxiliar o profissional da área de melhoramento genético animal com suas análises de dados por meio do incremento de qualidade de seus dados e da economia de tempo na preparação dos dados para entradas nos softwares mais populares.

Portanto, tem-se como hipótese que é possível se desenvolver um pacote para o sistema R que contenha funções para as tarefas mais comuns dentro de um processo de análise de dados genéticos.

2 OBJETIVOS

O objetivo do trabalho é dar início a um pacote de R com ferramentas que possam ser utilizadas por alunos e profissionais da área de melhoramento genético animal.

2.1 OBJETIVOS ESPECÍFICOS

Os objetivos específicos são:

1. Criar funções para análises de consistência de dados fenotípicos;

2. Criar funções para formatação de dados para o software de análise genética Wombat;
3. Escrever a documentação do pacote em português.

3 REVISÃO DE LITERATURA

3.1 Software R e seus pacotes

A utilização de ferramentas computacionais trouxe benefícios para a realização e execução de tarefas, otimizando o tempo e minimizando os erros no processo. Atualmente há inúmeras linguagens de programação que podem ser utilizadas.

O R é uma linguagem de programação *open source* para análise de dados, possui diversas ferramentas estatísticas e gráficas, pode ser utilizado sem ter custos de licença, atendendo estudantes e universidades, possui versões para Windows, GNU/Linux, Unix e MacOS (WALTER *et al.*, 2012).

O R foi desenvolvido em 1996 na Universidade de Auckland pelos pesquisadores Ross Ihaka e Robert Gentleman (PETERNELLI; MELO, 2011). Existem *interfaces* que são desenvolvidas por programadores que utilizam o R com o intuito de facilitar a compreensão dos usuários, de acordo com a sua necessidade, como por exemplo, o RExcel (BAIER; NEUWIRTH, 2007), Rcmdr (FOX, 2005) e o RStudio (RSTUDIO, 2022).

Uma das funcionalidades mais exploradas no R é a criação de funções que podem ser compiladas em pacotes, os pacotes oferecem ao usuário ferramentas para a realização de atividades que poupe o seu tempo. Alguns pacotes básicos são incluídos durante a instalação do R, entretanto, há outros pacotes disponíveis para serem acessados e instalados, atualmente há cerca de 18.000 pacotes na Rede Abrangente de Arquivos do R ou CRAN.

Por ser um software de código aberto, o usuário tem a liberdade de criar pacotes, e posteriormente, distribuí-lo para a comunidade gratuitamente, um dos motivos do R ser amplamente utilizado é a contribuição mútua entre usuários, assim sendo, é um software que está em constante evolução (VERZANI, 2008).

De acordo com Wickham e Bryan (2022), tudo que pode ser automatizado, deve ser automatizado, fazendo o mínimo de esforço com as mãos, e o máximo com as funções, muitos desenvolvedores têm a filosofia de que o tempo para a execução de uma determinada tarefa deve ser gasto na elaboração de um pacote, que posteriormente agilize a resolução do seu alvo de estudo ou resolução de um problema.

3.2 FERRAMENTAS PARA DESENVOLVIMENTO DE PACOTES EM R

A criação de um pacote pode ser feita de duas formas, a primeira é a maneira tradicional, utilizando comandos padrões do próprio R, como o R CMD *build*, usado para a criação e o R CMD *check*, que testa o pacote, e verifica se há algum problema no código.

Ambos são executados no terminal do programa, fora de uma sessão do R (WICKHAM; BRYAN, 2022). Utilizar apenas os comandos do próprio R para a criação e verificação de um pacote tem como vantagem a independência do usuário em não se prender a ferramentas adicionais além das que o R oferece, entretanto, estes comandos precisam ser executados fora da sessão do R, no terminal, fazendo com que a tarefa seja mais trabalhosa e demorada.

A segunda maneira é utilizar pacotes que nos auxiliam com funções para criar e conferir o pacote, testando e documentando as funções dentro de uma sessão do próprio R, onde o pacote está sendo desenvolvido, dois pacotes que permitem ao usuário aprimorar o desenvolvimento de seus estudos no R são o *devtools* (WICKHAM; CHANG, 2013) e *roxygen2* (WICKHAM; DANENBERG; EUGSTER, 2013), ambos podem ser carregados para uso por meio do comando *library()*.

Segundo Montgomery e Moore (2014) o pacote *devtools* poupa o trabalho do usuário em etapas de verificação e compilação do pacote no prompt do terminal do R, além de construir e checar as funções desenvolvidas, sendo possível simular a instalação do pacote dentro da sessão do R, tornando o fluxo de trabalho aprimorado.

Outro pacote que irá apoiar o usuário e facilitar a criação de pacotes é o *roxygen2* que atua formatando a documentação, que é um dos aspectos mais importantes para um bom código, a documentação do pacote expõe ao usuário o que está sendo feito, ou o que já foi feito no pacote, evitando que o usuário se perca no processo de formatação ou análise das funções presentes no pacote (WICKHAM; DANENBERG; EUGSTER, 2013).

3.3 PACOTES PARA MELHORAMENTO GENÉTICO ANIMAL EM R

De acordo com o CRAN do R existem alguns pacotes na área do melhoramento genético animal, que possuem algumas diferenças que propiciam ao usuário um leque de opções.

O STGS (NEERAJ *et al.*, 2019) foi desenvolvido para avaliar o mérito genético de um indivíduo através de marcadores do genoma.

O AGHMatrix (AMADEU *et al.*, 2016) constrói matriz de pedigree e/ou matriz (matriz A) de marcadores moleculares (matriz G), dando ao usuário a possibilidade de construir uma matriz combinada de pedigree corrigida por uma molecular (matriz H), trabalhando com dados diplóides e autopoliploides.

Outro pacote é QGglmm (DE VILLEMEREUIL, 2018) que fornece ao usuário funções para o cálculo da média, variância e herdabilidades dos dados observados.

O pacote gggroups (NILFOROOSHAN, M; SAAVEDRA-JIMÉNEZ, L., 2020) calcula matrizes com valores genéticos aditivos e de dominância.

O pacote optiSel (WELLMANN, 2019) possui um conjunto de funções para otimizar a seleção de animais com características desejáveis de acordo com os dados fornecidos pelo usuário.

O pacote sommer (COVARRUBIAS-PAZARAN, 2016) auxilia o usuário na seleção genômica.

O purgeR (LÓPEZ-CORTEGANO, 2022) é um pacote que possui funções de auxílio para a detecção e quantificação de depressão endogâmica, e expurgo genético em populações com pedigree.

Outro pacote é o GHap (UTSUNOMIYA *et al.*, 2016) identifica haplótipos presente nos dados, pontuando-os.

O pacote é o SFSI (LOPEZ-CRUZ; DE LOS CAMPOS, 2021) que avalia os dados para predição de valor genético baseado em índice de seleção.

O pacote xbreed (ESFANDYARI; SØRENSEN, 2019) faz simulações de populações com dados genômicos de animais de raça pura ou mestiços.

O pacote MDSMap (PREEDY; HACKETT; BOSKAMP, 2018) realiza um mapeamento de ligação genética para marcadores que estão presentes em um único cromossomo.

O pacote hip-hop (VAN DE POL; BROUWER; COCKBURN, 2022) pode ser utilizado para comparar os genótipos da prole com potenciais pais, através de marcadores genéticos bialélicos.

O pacote Cascade (JUNG *et al.*, 2014) oferece ferramentas para seleção de genes, engenharia reversa e previsão de dados em redes de cascata.

O pacote phyclust (CHEN, 2010) agrupa indivíduos a partir de dados moleculares, agrupando-os em subpopulações.

O quadro abaixo apresenta os pacotes existentes em comparação com o pacote proposto no trabalho, o LZ13.

Quadro 1 – Comparação entre pacotes existentes para R na área de melhoramento genético animal e o pacote proposto no presente trabalho

Identificação do pacote	Idioma da documentação	Consistência de dados fenotípicos	Consistência de dados de pedigree	Controle de qualidade de dados genômicos	Formatação dos dados para programas externos	Formação de grupos de contemporâneos
LZ13	Português/Inglês	☑	☑	☑	☑	☑
Stgs	Inglês	☑	☑	☑	☒	☒
AGHMatrix	Inglês	☑	☑	☑	☑	☒
QGglimm	Inglês	☑	☑	☑	☒	☒
gggroup	Inglês	☑	☑	☑	☒	☒
optiSel	Inglês	☑	☑	☑	☑	☒
sommer	Inglês	☑	☑	☑	☑	☒
purgeR	Inglês	☑	☑	☑	☒	☒
Ghap	Inglês	☑	☑	☒	☑	☒
SFSI	Inglês	☑	☑	☒	☑	☒
xbreed	Inglês	☑	☑	☑	☒	☒
MDSMap	Inglês	☑	☑	☒	☒	☒
hiphop	Inglês	☑	☑	☒	☒	☒
cascade	Inglês	☑	☑	☒	☒	☒
phyclust	Inglês	☑	☑	☒	☒	☒

Fonte: Elaborado pelo autor

4 MATERIAL E MÉTODOS

Material

Para dar suporte ao desenvolvimento, o Laboratório de Computação Científica Aplicada à Zootecnia - LuCCA-Z, da Universidade Estadual Paulista “Júlio de Mesquita Filho” - FCAT Faculdade de Ciências Agrárias e Tecnológicas - Campus de Dracena, dispõe de dois servidores: um Intel Core i7 3930K CPU 3.2Ghz, com 12 núcleos, 64GB de memória RAM e HD de 2TB; o segundo Intel Core i7 4930K CPU 3.4Ghz, com 12 núcleos, 48GB de memória RAM e HD 2TB, sendo acessados remotamente por SSH, interna ou externamente à unidade, utilizando chaves públicas e privadas. Ambos os servidores, funcionam com sistema operacional Linux Ubuntu

Server versão 20.04.2 LTS. Além dos servidores, foi utilizado um notebook de uso pessoal que auxiliou na construção das funções e nos testes.

Todo trabalho de desenvolvimento foi realizado utilizando o software R (R CORE TEAM, 2022) conjuntamente com a interface RStudio (RSTUDIO TEAM, 2022).

Para a realização dos testes das funções desenvolvidas utilizou-se bancos de dados reais sob nossa custódia. Os dados estão em formato original, contém todos os erros e imperfeições registrados nos momentos de coleta e genotipagem.

Métodos

Para o desenvolvimento das funções, buscou-se reproduzir o fluxo de trabalho utilizado no cotidiano da análise de dados; e simultaneamente evitar a utilização de outros pacotes que não o base do R, assim, minimizaríamos a dependência, de nossas funções, de terceiros. Nenhuma outra linguagem de programação, como por exemplo, C++, foi utilizada.

O desenvolvimento do nosso pacote e a documentação necessária serão criados e estarão de acordo com as normas e exigências dos pacotes *devtools* e *roxygen2*.

5 DESENVOLVIMENTO

A estimação de parâmetros genéticos como, por exemplo, a herdabilidade e valores genéticos aditivos (ou DEPs e PTAs) são parte importante dentro de programas de melhoramento genético para definição de estratégias e seleção de indivíduos. Nesses processos, utilizam-se dados fenotípicos, de pedigree e dados genômicos. Entretanto, para que os resultados sejam robustos e confiáveis é necessário que os dados utilizados passem por um processo de análise de consistência, uma vez que a coleta de dados no campo é muito propensa a erros de anotação/digitação pelo ritmo de trabalho e pelas condições/estrutura do local.

Para a consistência de dados fenotípicos, alguns procedimentos usualmente utilizados são:

1. Cálculo de estatísticas descritivas como média, moda, mediana, máximo e mínimo;
2. Geração de um histograma;
3. Geração de tabelas de frequência dos valores fenotípicos;
4. Cálculo de intervalos com base em desvio-padrão da característica;
5. Seleção/limpeza dos dados.

Média

Para avaliação de um banco de dados é importante a compreensão sobre como estes dados estão se comportando, a média apresenta uma medida sensível dos valores da amostra, e se estes valores são distribuídos de maneira uniforme, auxiliando o usuário em avaliações (BATANERO, 2000).

Moda

A moda de um conjunto de dados é o valor que ocorre com maior frequência, para análises de banco de dados na área do melhoramento genético animal é uma informação importante, já que mostra o número mais comum que ocorre nos valores estudados (BATANERO, 2000).

Mediana

A mediana em um conjunto de dados ordenados indica o valor que está exatamente no meio deste conjunto, ou seja, indica que metade dos valores estão acima e a outra metade estão abaixo deste valor, no contexto do melhoramento genético animal, a mediana pode auxiliar o usuário no entendimento de como estes valores estão se comportando, se a mediana é coerente com o foco do estudo (BATANERO, 2000).

Valores máximos e mínimos

Os valores máximos e mínimos indicam ao usuário como os dados estão distribuídos, podendo usar um comparativo em relação à média, se os valores estão muito distantes ou próximo da média, sendo importantes para avaliação de consistência dos dados (ARA; MUSETTI; SCHNEIDERMAN, 2003).

Histograma

Os histogramas auxiliam o usuário na identificação das distribuições dos dados, comunicando as informações graficamente, de imediato mostra ao usuário como os dados estão organizados, dividindo-os em classes (ARA; MUSETTI; SCHNEIDERMAN, 2003).

Tabela de frequência

As tabelas de frequência de frequência auxiliam o usuário na identificação rápida dos dados que foram pré-definidos em classes, facilitando a identificação de valores de interesse para o usuário (ARA; MUSETTI; SCHNEIDERMAN, 2003).

Desvio padrão

O desvio padrão é uma medida do nível de dispersão dos dados, indicando a uniformidade deste conjunto de dados, sendo uma medida importante para a tomada de decisão pelo usuário. Através do desvio padrão e contextualizando com as informações de medidas de tendências centrais, o usuário possui ferramentas para definir o que deverá ser feito (ARA; MUSETTI; SCHNEIDERMAN, 2003).

Seleção e limpeza dos dados

Para uma análise confiável, o usuário deve ser atentar em abastecer o R com um banco de dados sem inconsistências, ou seja, espaços em branco nas células da planilha ou tabela, caracteres que não fazem parte do contexto das informações como letras em posições dos números, espaços anormais entre valores, entre outros, esta etapa é fundamental para que o banco de dados esteja consistente para gerar resultados coerentes com o desejável pelo usuário (WANSBEEK; MEIJER, 2008).

Para cálculo da média, mediana e valores máximos e mínimos podemos utilizar a função *summary*. Os valores da média e mediana nos fornecem um valor que pode ser comparado com valores da literatura científica ou mesmo para avaliação com base na própria experiência do usuário. Valores da média muito acima ou abaixo do esperado pode ser um indicativo de dados com valores muito altos ou baixos (outliers).

O histograma pode ser gerado pela função *hist*. A finalidade é dar uma ideia visual do formato da distribuição dos dados.

A avaliação da média, mediana e moda, em conjunto, nos dão uma ideia sobre a simetria da distribuição. Para grande parte das características utilizadas em programas de melhoramento, espera-se uma distribuição normal (simétrica).

Para a moda, foi desenvolvida uma função própria, uma vez que não existe essa função no módulo base do R. Para calcular a moda o racional foi:

- O primeiro passo foi definir o nome que receberá a função, de maneira intuitiva, o nome foi “moda”;
- Criando o objeto m e atribuindo a ele a "table" x, x será fornecido pelo usuário, pode ser uma coluna de um banco de dados;
- Uso da função “*max*” para retornar à localização do valor máximo de m no objeto freqmax;
- Atribuindo a i valores de m contidos em freqmax através do operador *%in%*;
- Atribuindo m[i] à moda;

- Retornando à moda, ou seja, o valor que ocorre com maior frequência.

Abaixo o script:

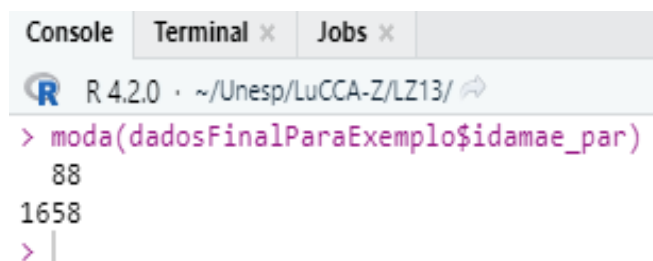
Figura 1 – Função moda

```
moda<-function(x){
  m<-table(x)
  freqmax<-max(m)
  i<-m%in%freqmax
  moda<-m[i]
  return(moda)
}
```

Fonte: Elaborado pelo autor

A seguir um exemplo da função utilizando a variável `idmae_par` (idade da mãe no parto) no banco de dados intitulado “dadosFinaisParaExemplo”.

Figura 2 – Exemplo da moda



```
Console Terminal x Jobs x
R 4.2.0 · ~/Unesp/LuCCA-Z/LZ13/
> moda(dadosFinalParaExemplo$idmae_par)
 88
1658
> |
```

Fonte: Elaborado pelo autor

O cálculo das tabelas de frequência pode ser realizado facilmente por meio da função `table`.

Se a distribuição é simétrica, particularmente normal, espera-se que mais do que 90% dos dados estejam dentro do intervalo (média $\pm$ desvio-padrão).

Durante cada passo, ou ao final desses passos, a seleção dos dados com erros e/ou de qualidade duvidosa devem ser eliminados. Após a eliminação, esses passos devem ser refeitos para verificação das mudanças causadas pela limpeza realizada. Dessa forma o script da função para o R é:

- Criando a função “consistF” referente a consistência de dados fenotípicos, `x` sendo o banco de dados do usuário e `ndp` o número de desvios padrão;
- Utilizando o comando `summary` para resultados como média, mediana, valores máximos e mínimos, antes e depois da limpeza dos dados;

- Criação de histograma com a função *hist* para antes da consistência dos dados;
- Criando parâmetros para simulação de eliminação de outliers através da função de desvio padrão (*sd*), indicando o grau de variação dos elementos;
- Eliminando os dados de *x* para uma nova análise
- Criação de histograma com a função *hist* depois da consistência dos dados. Abaixo o script:

Figura 3 – Função de consistência dos dados

```
consistF<-function(x,ndp=3){
  #Estatísticas descritivas
  estatDesc<-list(resumoAntes=NULL,freqAntes=NULL,resumoDepois=NULL,freqDepois=NULL)
  estatDesc$resumoAntes<-summary(x)
  estatDesc$freqAntes<-table(x)

  #Ajustando parâmetros gráficos
  oldpar<-par(mfrow=c(1,2))

  #Histograma
  hist(x,col="red",main="Antes da consistência",xlab="y",ylab="Frequência")

  #Parâmetros para eliminação de outliers
  dp<-sd(x)
  media<-mean(x)
  linf<-media-(ndp*dp)
  lsup<-media+(ndp*dp)
  print(paste("Limite inferior para a eliminação de dados: ",round(linf,2),sep=""))
  print(paste("Limite superior para a eliminação de dados: ",round(lsup,2),sep=""))

  #Simulação da eliminação de outliers
  elimlinf<-sum(x<linf)
  print(paste("Dados eliminados pelo limite inferior: ",elimlinf))
  elimlsup<-sum(x>lsup)
  print(paste("Dados eliminados pelo limite superior: ",elimlsup))
  elimTotal<-elimlinf+elimlsup
  print(paste("Total de dados eliminados: ",elimTotal))

  #Eliminação dos dados para cálculo das novas estatísticas descritivas
  x[x<linf]<-NA
  x[x>lsup]<-NA

  #Estatísticas descritivas após a eliminação dos dados
  estatDesc$resumoDepois<-summary(x)
  estatDesc$freqDepois<-table(x)

  #Histograma após a eliminação dos dados
  hist(x,col="green",main="Depois da consistência",xlab="y",ylab="Frequência")

  return(estatDesc)
}
```

Fonte: Elaborado pelo autor

A seguir um exemplo da função `consistF` para análise de consistência fenotípica, utilizando a variável `idamae_par` (idade da mãe no parto) do banco de dados intitulado “`dadosFinalParaExemplo`”.

Figura 4 – Exemplo da função `consistF`

```

Console Terminal x Jobs x
R 4.2.0 · ~/Unesp/LuCCA-Z/LZ13/

Restarting R session...

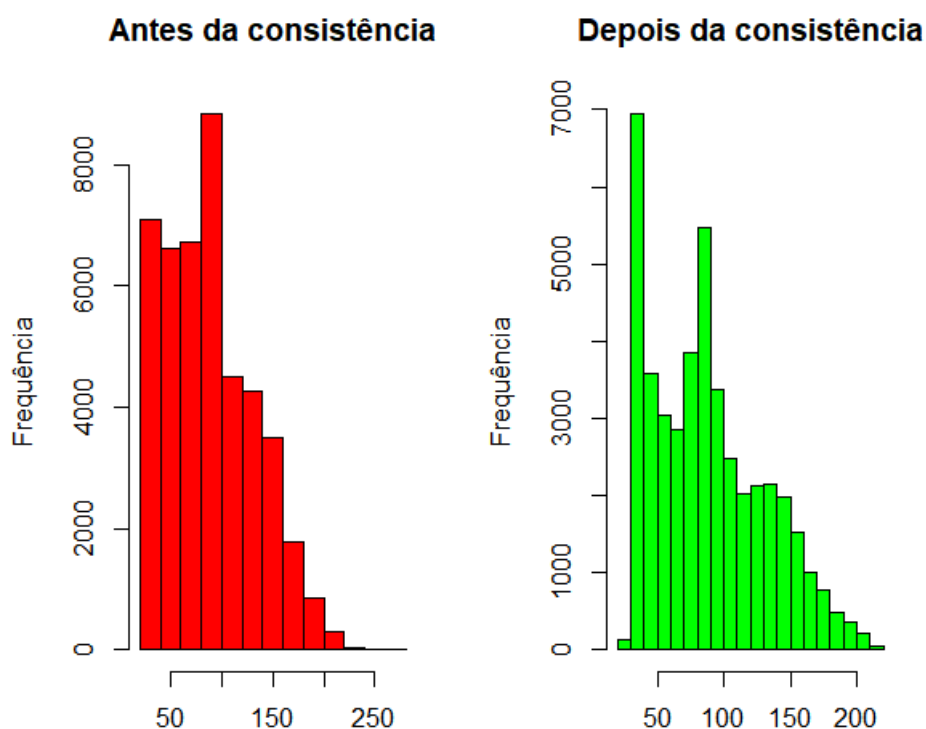
> library(LZ13)
> consistF(dadosFinalParaExemplo$idamae_par)
[1] "Limite inferior para a eliminação de dados: -38.43"
[1] "Limite superior para a eliminação de dados: 215.91"
[1] "Dados eliminados pelo limite inferior: 0"
[1] "Dados eliminados pelo limite superior: 76"
[1] "Total de dados eliminados: 76"

```

Fonte: Elaborado pelo autor

Outro exemplo são os gráficos de frequências geradas pela função para que o usuário analise de maneira visual os dados:

Figura 5 – Gráficos de frequências da função `consistF` para a variável `idamae_par`



Fonte: Elaborado pelo autor

Outro processo importante na estimação de parâmetros genéticos e predição de valores genéticos aditivos é a formação de grupos de contemporâneos (GC). O

efeito dos GC representa um efeito de ambiente global e comum a vários indivíduos da população que pode ser corrigido no processo de estimação/predição. A ideia é atribuir um código a todo indivíduo que foi criado sob as mesmas condições de manejo e que nasceram no mesmo ano e estação (portanto, tiveram a mesma qualidade e disponibilidade de alimentos e os mesmos desafios com relação a parasitas e doenças). A função foi desenvolvida seguindo alguns passos:

- Definindo o nome da função como `gc`, a qual recebe como argumentos as colunas que devem ser combinadas para a formação dos grupos de contemporâneos(...) o número de indivíduos dentro de cada grupo, definido por `n` e `criar.gc=FALSE`, que define se o grupo deve ser criado no *data frame*. O padrão `FALSE` tem como finalidade permitir que o usuário faça estudos preliminares dos grupos sem escrevê-los definitivamente nos dados. Assim que o `gc` estiver em condições de ser criado, o usuário irá substituir o `criar.gc=FALSE` para `criar.gc=TRUE`;
- Formação de uma coluna de grupo de contemporâneos, atribuindo o nome de `gctemp` com o `NULL`, ou seja, `gctemp` é temporário e receberá valores quando o R fizer a execução de um *loop* para percorrer os argumentos encontrados em (...), os quais serão concatenados pela função `paste0` e incluídos em `l` (uma lista);
- A simulação dos dados permite ao usuário avaliar se o valor `n` indicado no argumento não vai gerar uma grande perda de informações para a análise;

Abaixo a função desenvolvida:

Figura 6 – Função para grupo de contemporâneos

```
gc<-function(...,data,n,criar.gc=FALSE){
  l<-list(...)

  #Formando a coluna de grupos de contemporâneos
  gctemp<-NULL
  for(i in 1:length(...)){
    gctemp<-paste0(gctemp,l[[i]])
  }

  #Simulando a eliminação de dados
  freqGC<-table(gctemp)
  print(paste("Número total de grupos de contemporâneos formados: ", length(names(freqGC))))
  dadosTotal<-sum(freqGC)
  freqGC1<-freqGC[freqGC>=n]
  dadosMant<-sum(freqGC1)
  print(paste("Número total de grupos de contemporâneos após eliminação: ", length(names(freqGC1))))
  print(paste("Dados eliminados: ", dadosTotal-dadosMant,sep=" "))

  #Criando os grupos no data frame original
  if(criar.gc){
    data<-data.frame(data,gc=NA)
    i<-gctemp%in%names(freqGC1)
    data$gc[i]<-gctemp[i]
    data<-data[!is.na(data$gc),]
  }
  return(data)
}
```

Fonte: Elaborado pelo autor

A seguir um exemplo para a formação de grupo de contemporâneos, utilizando o banco de dados intitulado “dadosFinalParaExemplo (1)” utilizando os argumentos (variáveis) ano e mês.

Figura 7 – Exemplo da formação de grupo de contemporâneos

```
> gc(`dadosFinalParaExemplo (1)`$ano,`dadosFinalParaExemplo (1)`$mes,dadosFinalParaExemplo,4,criar.gc = FALSE)
[1] "Número total de grupos de contemporâneos formados: 2460"
```

Fonte: Elaborado pelo autor

Finalmente, com os dados consistentes, devemos utilizá-los nos programas específicos para estimação de componentes de variância e predição de valores genéticos aditivos.

Esses programas exigem um formato específico dos dados para que esses sejam lidos de forma correta. Nessa etapa, muitos problemas acontecem por alguma exigência não ter sido cumprida e tempo valioso é perdido até que se iniciem as análises. Sendo assim, uma função que auxiliasse na formatação dos dados garantiria que as exigências fossem cumpridas e bastante tempo economizado.

Inicialmente, a função desenvolvida foi para o software Wombat. Segundo Meyer (2007) as exigências para o arquivo de fenótipos são:

- Os arquivos podem ser de dados, pedigree, marcadores, entre outros;
- Definir o nome do arquivo de saída (em até 30 caracteres);
- Variáveis devem estar em colunas de larguras fixas separadas por espaço;
- Todas as colunas têm que ser preenchidas com valores numéricos (sem espaços em branco);
- Todos os códigos dos efeitos considerados devem ser inteiros positivos;
- Valor máximo permitido para os códigos é de 2147483647;
- Para análises multi-características, deve haver um registro para cada característica medida para um indivíduo;
- O número da característica para o registro deve ser fornecido na primeira coluna;
- Não está disponível código para dados perdidos. Dados perdidos são registros ausentes no arquivo de dados;
- O arquivo de dados deve ser ordenado em ordem ascendente, de acordo com:
 - i) o indivíduo, para o qual características foram registradas, e
 - ii) número da característica dentro do indivíduo
 - iii) Para análises de Regressão Aleatória, os registros devem ser ordenados de acordo com o valor da variável controle (dentro de indivíduo e número da característica);

- Para análises multivariadas que combinam registros simples e repetidos, as características com registros repetidos precisam ter um número de característica menor do que aquelas de registros simples;
- Para facilitar a anotação do arquivo de dados (cabeçalhos de colunas, data da criação, fonte), Wombat ignora linhas com uma "#" na coluna 1, no início do arquivo;
- Não existe limite para o número de linhas de comentários, mas elas devem ser as primeiras linhas.

A seguir o script:

Figura 8 – Função auxiliar de formatação dos dados para o Wombat

```
#formatar para Wombat
format_wombat<-function(df,file="formattedWombat.txt",type="data",
                        ndigit=rep(8,length(df)),missing=-999.999,varCol,codCol){
  if(type=="data"){
    #Inserindo valores perdidos
    i<-is.na(df)
    df[i]<-missing

    #Verificando se os dados são numéricos
    stopifnot(sapply(dados,is.numeric))

    #Verificando as escalas das variâncias
    stopifnot(!(diag(var(dados))<10e-5|diag(var(dados))>10e5))

    #formatando o arquivo
    formato<-NULL
    argum<-list()
    for(i in 1:length(df)){
      if(i==length(df)){
        formato<-paste0(formato,"%",".",ndigit[i],"g")
      }else{
        formato<-paste0(formato,"%",".",ndigit[i],"g ")
      }
      argum[i+1]<-df[i]
    }
    print(formato)
    argum[1]<-formato

    linhas<-do.call(sprintf,argum)
    writeLines(linhas,file)
  }
  if(type=="pedigree"){
    "dados de pedigree"
  }
  if(type=="genomic"){
    "dados genômicos"
  }
}
```

Fonte: Elaborado pelo autor

A seguir um exemplo da função `format_wombat` e os dados gerados, são estes dados que o usuário irá incluir no programa de avaliação genética Wombat.

Figura 9 – Exemplo da função Wombat

```
> format_wombat(dados)
-0.83271338 10.8579970.91560841
10.0050330.93082468 8.1932033-
0.013348997 11.1181910.19159944
13.788068
```

Fonte: Elaborado pelo autor

Finalmente para que o trabalho esteja completo é necessário empacotar essas funções de forma que possam ser utilizadas por terceiros e tenham sua utilização facilitada. Os pacotes servem para compilar as funções para que possam ser utilizadas pelo usuário e/ou disponibilizada para outros usuários via o repositório do R, permitindo que os pacotes possam ser aprimorados com novas funções, ampliando os objetos de estudos, contextualizando diferentes áreas do conhecimento, no melhoramento genético animal os usuários que trabalham com o R podem traçar novas estratégias, aprimorando as suas análises e tornando mais eficiente a tomada de decisão. Para criação do pacote o usuário deve ir na interface do R, no canto superior esquerdo, em *File* → *New Project* → *New Directory* → *R Package* → Em “*package name*” dê nome ao pacote → *Create Project*. Pronto, o pacote foi criado. Nesta segunda etapa o usuário deve ir até a aba *New File* → *R script*, no *R script* o usuário irá escrever as suas funções. Quando as funções estiverem prontas, é fundamental que este instale no R dois pacotes, o *devtools* e o *roxygen2* utilizando o comando `install.packages()`, ou seja `install.packages("devtools")` e `install.packages("roxygen2")`, ambos irão facilitar a documentação e a checagem do pacote. Após os passos anteriores, a próxima etapa é utilizar o comando `#' @export` na primeira linha do *R script*, acima das funções que estão sendo desenvolvidas ou aquelas já concluídas, este comando fará com que todas as funções feitas pelo usuário possam ser(em) exportada(s), e posteriormente, na seção de *Console* do R o usuário deverá apertar `CRTL+SHIFT+B` para que assim as funções fiquem disponíveis. Para utilizar as funções que já foram compiladas no pacote é fundamental definir a documentação, esta etapa pode ser feita com o comando `#' @param` que define os parâmetros da função e o comando `#' @examples` que recebe um exemplo didático informado pelo usuário, ambos os comandos devem ser incluídos no *Console* do R, acima das funções e do `#' @export`.

Para utilizar o pacote e as funções que já estão documentadas, o usuário irá acessar o pacote através do comando `library("nome do pacote")`, já com as funções documentadas, toda a descrição do pacote pode ser acessada na aba *Help*, ali estará as informações sobre a função que o usuário está utilizando. A descrição do pacote pode ser acessada pelo arquivo *DESCRIPTION*, a figura abaixo mostra as informações do pacote, ou seja, a sua descrição.

Figura 10 – Descrição do pacote LZ13

```

Package: LZ13
Type: Package
Title: Funções Auxiliares Para Análises de Dados Genéticos e Genômicos
Version: 0.1.0
Author: Who wrote it
Maintainer: The package maintainer <yourself@somewhere.net>
Description: More about what it does (maybe more than one line)
             Use four spaces when indenting paragraphs within the Description.
License: What license is it under?
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.2.3

```

Fonte: Elaborado pelo autor

Com todas as funções feitas e documentadas, o pacote LZ13 possui duas funções para análise de consistência de dados, uma para formação de grupo de contemporâneos e uma função auxiliar, para formatação de dados que poderão ser utilizados no Wombat, estas funções podem ser acessadas pelo arquivo *NAMESPACE*, como ilustrado na figura a seguir.

Figura 11 – Funções do pacote LZ13

```

# Generated by roxygen2: do not edit by hand

export(consistF)
export(format_wombat)
export(gc)
export(modas)

```

Fonte: Elaborado pelo autor

6 CONCLUSÃO

O pacote LZ13 é desenvolvido para auxiliar os usuários em análises de banco de dados na área do melhoramento genético animal, através de funções que permitem a limpeza de dados inconsistentes, a formação e análise de qualidade dos grupos de contemporâneos e a formatação dos dados para o uso em programas de avaliações genéticas.

7 REFERÊNCIAS

- AMADEU, R.; CELLON, C.; OLMSTEAD, J.; GARCIA, A.; RESENDE, M; MUÑOZ, P. **AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example**. The plant genome, v.9, n.3, p. 1-10, 2016.
- ARA, A.; MUSETTI, A.; SCHNEIDERMAN, B. **Introdução à estatística**. Editora Blucher, 1ª Edição, p. 3-34, 2003.
- BAIER, T.; NEUWIRTH, E. **Excel:: COM :: R. Computational Statistics**, v. 22, n. 1, p. 91- 108, 2007.
- BATANERO, C. **Significado y comprensión de las medidas de posición central**. Departamento de Didáctica de la Matemática, Universidad de Granada, p. 1-12, 2000.
- CHEN, W.-C. **A Quick Guide for the phyclust Package**. [s. l.], p. 1-26, 2010.
- COVARRUBIAS-PAZARAN, G. **Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer**. PLOS ONE, [s. l.], v. 11, n. 6, p. 1-15, 2016.
- DE VILLEMEREUIL, P. How to use the QGglmm package?. **Overview of the theory Quantitative genetics and linear mixed models**, p. 1-31, 2018.
- ESFANDYARI, H.; SØRENSEN, A. C. **xbreed: An R package for genomic simulation of purebred and crossbred populations**. [s. l.], p. 1-32, 2019.
- FOX, J. **The R Commander: A Basic-Statistics Graphical User Interface to R**. Journal of Statistical Software, v. 14, n. 9, p. 1-42, 2005.
- JUNG, N.; BERTRAND, F.; BAHRAM, S.; VALLAT, L.; MAUMY-BERTRAND, M. **Cascade: a R package to study, predict and simulate the diffusion of a signal through a temporal gene network**. Bioinformatics, [s. l.], v. 30, n. 4, p. 571–573, 2014.
- LÓPEZ-CORTEGANO, E. **purgeR: inbreeding and purging in pedigreed populations**. Bioinformatics, [s. l.], v. 38, n. 2, p. 564–565, 2022.
- LOPEZ-CRUZ, M.; DE LOS CAMPOS, G. **Optimal breeding-value prediction using a sparse selection index**. Genetics, [s. l.], v. 218, n. 1, p. 1-10, 2021.
- MEYER, K. Wombat – **A tool for mixed model analyses in quantitative genetics by REML**, j. Zhejiang Uni. SCIENCE B. p. 815–821, 2007.
- MONTGOMERY, J.; MOORE, R. T. **Building and Maintaining R Packages with devtools and roxygen2** [s. l.], v. 22, p. 1-6, 2014.
- NEERAJ B.; DC M.; ANIL R.; KK C., **Package ‘STGS’. Molecular Breeding**, v. 1, p. 201-210, 2019.
- NILFOROOSHAN, M; SAAVEDRA-JIMÉNEZ, L. **Ggroups: an r package for pedigree and genetic groups data**. Hereditas, V. 157, N. 1, p. 1-9, 2020.

PETERNELLI, L. A.; MELLO, M. P. **Conhecendo o R: Uma Visão Estatística**. 1a ed. Viçosa: Editora UFV. p. 1-185, 2011.

PREEDY, F.; HACKETT, A.; BOSKAMP, B. **MDSMap: High density linkage maps using multi-dimensional scaling**. [s. l.], p. 1-10, 2018.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em: 30 nov. 2022.

RSTUDIO TEAM. **RStudio: Integrated Development Environment for R**. RStudio, PBC, Boston, MA. Disponível em: <http://www.rstudio.com/>. Acesso em: 30 nov. 2022.

UTSUNOMIYA, T.; MILANESI, M.; UTSUNOMIYA, H.; AJMONE-MARSAN, P.; GARCIA, F. **GHap: an R package for genome-wide haplotyping**. Bioinformatics, [s. l.], v. 32, n. 18, p. 2861–2862, 2016.

VAN DE POL, M.; BROUWER, L.; COCKBURN, A. **hiphop: Parentage Assignment using Bi-Allelic Genetic Markers**. Comprehensive R Archive Network (CRAN). Disponível em: <https://pure.knaw.nl/portal/en/publications/07cbf326-fb29-4b3d-9452-a7eb4c043b70>. Acesso em: 30 nov. 2022.

VERZANI, J. **Using R in Introductory Statistics Courses with the pmg Graphical User Interface**. Journal of Statistics Education, v. 16, n. 1, p. 01-17, 2008.

WALTER, O. M. F. C.; HENNING E.; KONRATH, C. A.; ALVES, C. C.; SAMOBYL, W. R. **Uma Visão Geral do RStudio Aplicado ao Ensino de Controle Estatístico do Processo**. In: XL CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA. p. 2-4, 2012.

WANSBEEK, T.; MEIJER E. **“Measurement Error and Latent Variables.” A Companion to Theoretical Econometrics**. John Wiley & Sons, p. 1-162, 2008.

WELLMANN, R. **Optimum contribution selection for animal breeding and conservation: the R package optiSel**. BMC Bioinformatics, [s. l.], v. 20, n. 1, p. 1-25, 2019.

WICKHAM, H.; BRYAN, J., **R Packages**, 2. ed., [s. l.]: O’Reilly Media. 2022.

WICKHAM, H.; CHANG, W., **devtools: Tools to make developing R code easier**. R package version 1.1. 2013.

WICKHAM, HADLEY; DANENBERG, PETER; EUGSTER, MANUEL. **roxygen2: In-source documentation for R**. R package version, v. 3, n. 0, 2013.