CRISTIANO FERREIRA DE OLIVEIRA

GOLIAS: SOFTWARE PARA ANÁLISES ESTATÍSTICA E BIOMÉTRICA DE GRANDE CONJUNTO DE DADOS UTILIZANDO AS LINGUAGENS JÚLIA E R

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

O48
2018

Oliveira, Cristiano Ferreira de, 1990-
Golias : software para análises estatística e biometrica de grande conjunto de dados utilizando as linguagens Júlia e R / Cristiano Ferreira de Oliveira. – Viçosa, MG, 2018.
viii, 48f. : il. ; 29 cm.

Orientador: Cosme Damião Cruz.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Biometria. 2. Golias (Software). 3. Melhoramento genético - Software. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 570.15195

CRISTIANO FERREIRA DE OLIVEIRA

GOLIAS: SOFTWARE PARA ANÁLISES ESTATÍSTICA E BIOMÉTRICA DE GRANDE CONJUNTO DE DADOS USANDO AS LINGUAGENS JÚLIA E R

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 21 de fevereiro de 2018.

Leonardo Lopes Bhering

Moysés Nascimento
(Coorientador)

Cosme Damião Cruz
(Orientador)

“Fé não faz as coisas serem fáceis, mas as tornam totalmente possíveis, creia. “

Romanos 5:8

“Não se emolde ao padrão deste mundo, mas transforme-se!”

Romanos 12:1

“A imaginação é mais importante que o conhecimento. O conhecimento é limitado, enquanto a imaginação abraça o mundo inteiro, estimulando o progresso, dando à luz à evolução.”

Albert Einstein

AGRADECIMENTOS

Ao meu pai, Luiz e à minha mãe Ana pelo carinho e dedicação. Aos meus irmãos pela amizade incondicional.

À Universidade Federal de Viçosa, particularmente ao Departamento de Estatística que contribuíram enormemente para minha formação profissional e pessoal.

Aos amigos funcionários e professores da UFV que sempre nos apoiam.

Ao meu orientador, Cosme Damião Cruz, pelos conselhos, amizade, paciência e incontestável orientação.

Ao professor e co-orientador Moysés pelas valiosas contribuições.

Aos amigos do Laboratório de Bioinformática, pela amizade e companheirismo.

À minha família e também aos amigos que nestes anos se tornaram família.

À CAPES pelo apoio financeiro para o desenvolvimento deste trabalho.

BIOGRAFIA

CRISTIANO FERREIRA DE OLIVEIRA, filho de Luiz Lourenço de Oliveira e Ana Maria Batista de Oliveira, nascido em 27 de dezembro de 1990, em Paula Cândido no estado de Minas Gerais.

Cursou o ensino fundamental e médio na rede pública na cidade de Paula Cândido/MG. Em 2010 iniciou o curso de Licenciatura em Matemática, na Universidade Federal de Viçosa -UFV, graduando em janeiro de 2016.

Em fevereiro de 2016 iniciou o Mestrado *Stricto Sensu* no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 21 de fevereiro de 2018.

SUMÁRIO

RESUMO vii
ABSTRACT viii
1 Introdução Geral 1
2 Revisão de Literatura 5
2.1 Marcadores SNP: conceitos básicos 5
2.2 Análise genética com grandes bancos de dados 5
2.3 Importância de alguns estudos genéticos 6
a) Estrutura da população 6
b) Análise genética em grandes conjuntos de dados 6
2.4 Descrição da estrutura genética de uma população 7
2.4.1 Frequências alélicas e genotípicas de uma população 7
2.4.2 Equilíbrio de Hardy-Weinberg 7
2.5 Análise da dimensionalidade por Componentes Principais 9
3 Bibliografia 12
CAPÍTULO 01 14
RESUMO 15
ABSTRACT 16
1 Introdução 17
2 Material e Métodos 19
I. Descarte de variáveis via Análise de Componentes Principais (PCA) 19
II. Controle de qualidade de SNP's 19
a) Call Rate 19
b) Frequência do alelo mais raro (MAF) 20
c) Equilíbrio de Hardy-Weinberg 20
2.1.1 Imputação de dados genotípicos 20
III. Modelos utilizados para Seleção Genômica 21
Utilizando o Golias 21
Exemplo de aplicação: 22
2.1.2 Call Rate 22
2.1.3 MAF 23
2.1.4 Imputação dos dados Genotípicos 24
3 Resultados e Discussão 24
4 Agradecimentos 24
5 Referências Bibliográficas 25
CAPÍTULO 02 26
RESUMO 27
ABSTRACT 28

1 Introdução 29
2 Material e Métodos 31
2.1 Material 31
2.2 Metodologias 32
3 Resultado e Discussão 37
3.1 Filtragem no banco de dados moleculares; 37
3.2 Comparação de desempenho do pacote BLGR nos ambientes R e Júlia ... 37
4 Conclusão 46
5 Referências Bibliográficas 47

RESUMO

Oliveira, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Golias: Software para análises estatística e biométrica de grande conjunto de dados utilizando as linguagens Júlia e R.** Orientador: Cosme Damião Cruz. Coorientador: Moysés Nascimento.

A utilização de informações de marcadores moleculares para identificação de indivíduos geneticamente superiores para características de interesse é uma as principais contribuições da genética molecular no melhoramento genético. Assim, em programas de melhoramento genético, é crescente o uso de marcadores moleculares dispostos em matrizes de alta densidade, que necessitam de controle de qualidade e uso de metodologias adequadas para realizar a predição dos valores genéticos. Diante do exposto este trabalho tem como objetivo apresentar o software Golias, que é um software livre destinado ao processamento de dados de marcadores SNPs de alta dimensão. O Golias é integrado com os ambientes de programação R e Julia, utilizando o poder de processamento da linguagem Júlia e de pacotes do R otimizados em código C. Todas as análises são realizadas sem fazer necessário que o usuário tenha qualquer conhecimento em programação. O software proposto dispõe de procedimentos para o controle de qualidade de SNPs como call rate, MAF, teste de equilíbrio de Hardy-Weinberg, descarte de variáveis via componentes principais, um procedimento para imputação de informações genotípicas perdidas e algumas técnicas de predição utilizando métodos bayesianos. O Golias está disponível em português; pode ser baixado da Internet (https://www.dropbox.com/sh/ql92mzjed8735wq/AADfVadXY3oYM_3qe84VbN-Ea?dl=0) e é compatível com sistema operacional Windows.

ABSTRACT

OLIVEIRA, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, February, 2018. **Golias: Software for statistical and biometric analysis of a large dataset using the languages Julia and R.** Adviser: Cosme Damião Cruz. Co-Adviser: Moysés Nascimento.

The use of molecular marker information to identify genetically superior individuals for traits of interest is one of the main contributions of genomic selection in breeding. Thus, in genetic breeding programs is increasing the use of molecular markers arranged in high density matrices which require quality control and the use of appropriate methodologies to predict genetic values. The work aims to present the Golias software that is a free software for processing data from high-size SNPs. Golias is integrated with the R and Julia programming environments, using the processing power of the Júlia language and R-optimized C-packets. All analyzes are performed without making it necessary for the user to have any programming knowledge. The proposed software has procedures for quality control of SNPs such as call rate, MAF, Hardy-Weinberg equilibrium test, variables discarding via principal components, a procedure for imputation of lost genotype information and some prediction techniques using Bayesian methods. Golias is available in Portuguese; can be downloaded from the Internet (https://www.dropbox.com/sh/ql92mzjed8735wq/AADfVadXY3oYM_3qe84VbN-Ea?dl=0) and is compatible with Windows operating system.

1 Introdução Geral

A redução no custo do processo de genotipagem permitiu a produção de um número massivo de dados genômicos. Em consequência, surge grande mudança de paradigmas, sobretudo naqueles temas que permeiam as dimensões dos dados (RESENDE; SILVA; RESENDE JÚNIOR, 2013).

Na área de ciências agrárias, e mais especificamente nos programas de melhoramento genético, pode-se observar a evolução nas técnicas de determinação de marcas moleculares, conhecido por genotipagem, que contribuem para tomadas de decisão e representam uma grande informação sobre o indivíduo a partir de seu DNA. O desenvolvimento das técnicas de eletroforese de isoenzimas revolucionou os estudos em genética durante a década de 95 (CRUZ; MEDEIROS, FÁBIO FERREIRA PESSONI, 2011). Porém uma das principais limitações dos marcadores isoenzimáticos é a reduzida cobertura nos genomas estudados, devido ao pequeno número de locos que podem ser detectados (CAIXETA; BORÉM, 2006).

Posteriormente, surgiram as técnicas de biologia molecular dando origem a métodos de detecção de polimorfismo em nível de DNA como por exemplo os marcadores RFLP (restrict fragment length polymorphism), RAPD (random amplified polymorphic DNA), AFLP (amplified fragment length polymorphism) e SNPs (single nucleotide polymorphisms) e SSCP-SNP (Marcador molecular co-dominante e robusto baseado em SNPs) que permitem caracterizar grande número de locos.

A utilização de informações de marcadores moleculares se destaca como uma as principais contribuições da genética molecular no melhoramento genético (AZEVEDO, 2012). Agregando informações fenotípicas às informações moleculares é possível predizer valores genéticos de forma mais acurada. Neste contexto, a abordagem por meio de modelos de seleção genômica ampla (GWS) se destaca, pois em seus modelos contemplam as informações de grandes números de marcadores amplamente distribuídos ao longo do genoma.

O uso de conjuntos de dados de SNP’s de alta densidade é cada vez mais comum no melhoramento genético. Estes conjuntos possuem um número de marcadores extremamente maior que o número de indivíduos genotipados e, assim,

apesar de proporcionar ampla informação sobre o indivíduo trazem para dentro de procedimentos biométricos questões importantes como o controle de qualidade dos dados e também a necessidade de métodos não convencionais para realizar as análises.

Métodos adequados de controle e garantia de qualidade dos dados possuem grande importância pois podem evitar resultados espúrios (LAURIE et al., 2010). Assim, como a maioria dos procedimentos de predição de valores genéticos consiste em ajustar modelos em que os marcadores moleculares são as variáveis explicativas, tem-se vários desafios estatísticos e computacionais.

Em termos de desafios computacionais tem-se a demanda de hardwares e softwares, capazes de processar quantidades massivas de dados. No âmbito estatístico surge uma problemática quando o número de marcadores iguala ou supera o número de indivíduos genotipados. Isto pode causar problemas de identificação paramétrica devido a colinearidade das variáveis preditoras ou então a inviabilidade dos estimadores de quadrados mínimos. Além disto, devido a dimensão dos dados, na GWS o número de graus de liberdade para estimar os efeitos de todos os marcadores torna-se escasso (RESENDE; SILVA; AZEVEDO, 2014).

Soluções para contornar problemas tem sido abordado por meio de métodos Bayesianos (PÉREZ; DE LOS CAMPOS, 2014) como BayesA e BayesB, métodos fundamentados em inteligência computacional tal como redes neuras e métodos de regressão com redução de dimensionalidade como o de componentes principais (RESENDE; SILVA; AZEVEDO, 2014).

Para realizar o processamento destes dados, tem sido disponibilizado vários softwares gerados a partir de várias linguagens de programação tais como C++, R, Python, Java, Julia, MATLAB, dentre outros. Porém nem sempre é fácil decidir por qual optar, uma vez que para trabalhar com quantidades massivas de dados o desempenho do aplicativo na linguagem utilizada passa a ser altamente relevante pois influencia diretamente no tempo gasto para realizar o processo.

Isto pode ser notado na figura 1, onde é apresentado os resultados de um benchmark (“Julia Micro-Benchmarks”, 2018), no qual é comparada o tempo de processamento para oito funções em diferentes linguagens.

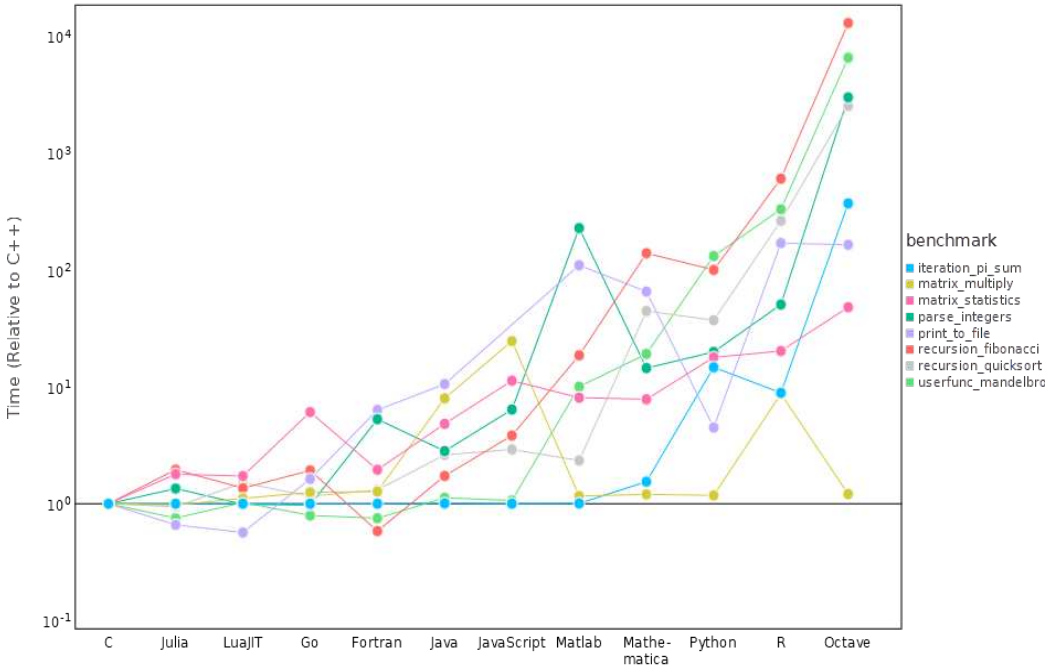


Figura 1. Tempo de execução em relação ao obtido com a linguagem C (desempenho de C = 1,0).

O ambiente de programação R (https://www.r-project.org/) possui grande destaque na comunidade acadêmica. De maneira geral, o R é uma linguagem destinada à computação estatística e geração gráficos em análises estatísticas. Ele dispõe de ferramentas estatísticas tais como modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, dentre outros. Para tarefas computacionais que exigem maior desempenho, destacam-se as linguagens C, C ++ e o código Fortran. Scripts desenvolvidos nesta linguagens podem ser vinculados ao R e chamados, em tempo de execução, permitindo que usuários possam escrever código, em linguagens externas, para manipular objetos diretamente no R.

Outra linguagem que também vem ganhando espaço é a linguagem Julia (http://julialang.org/), com uma proposta de fácil sintaxe e alto poder de processamento. Projetada para alcançar alto desempenho em computação numérica e científica, Julia é vista como uma alternativa de alto desempenho, fácil sintaxe e de código aberto para linguagens como MATLAB, R e Python. (PINTOS et al., 2016).

Desta forma este trabalho tem como objetivo apresentar o software Golias. O Golias é um software livre destinado ao processamento de conjuntos de SNPs de

alta dimensão. É compatível e integrado com os ambientes de programação R e Julia ambos gratuitos. O Golias utiliza o poder de processamento da linguagem Julia e de pacotes do R otimizados em código C. Ele permite ao usuário realizar todas as análises sem ter qualquer conhecimento em programação.

Em sua primeira versão, estão disponíveis procedimentos voltados para o controle de qualidade de marcadores moleculares SNPs como call rate, MAF, teste de equilíbrio de Hardy-Weinberg, descarte de variáveis via componentes principais e algumas técnicas de modelagem utilizando métodos bayesianos.

2 Revisão de Literatura

2.1 Marcadores SNP: conceitos básicos

Os marcadores SNP tem como base mutações em bases únicas da cadeia de bases nitrogenadas (Adenina, Citosina, Timina e Guanina). Eles se destacam com relação a outros tipos de marcadores genéticos devido à sua baixa taxa de mutação, facilidade de genotipagem e baixo custo (RESENDE; SILVA; AZEVEDO, 2014).

Considere os indivíduos a seguir:

Indivíduo 1: TCA**C**GTGGC

Indivíduo 2: TCA**T**GTGGC

Pode-se verificar polimorfismo de SNPs entre eles, uma vez que na sequência especificada ocorre troca de uma de suas bases.

2.2 Análise genética com grandes bancos de dados

Nos últimos anos as tecnologias de sequenciamento de DNA passaram por vários ciclos de renovação. Houve considerável redução no custo do processo de genotipagem permitindo, assim, a produção de um número massivo de dados genômicos. Em consequência deste fato, surge grande mudança de paradigmas, sobretudo naqueles temas que permeiam a escala das análises (RESENDE; SILVA; RESENDE JÚNIOR, 2013). Deve ser considerado que os marcadores moleculares como SNP, além de proporcionar grande redução no custo do processo de genotipagem permite avaliação genética em nível molecular eficiente revelando ao pesquisador informações valiosas sobre indivíduos e populações. Chips de DNA de alta densidade foram criados para genotipar dezenas e até centenas de milhares de marcadores SNP em um único ensaio. Essas novas tecnologias permitiram a geração de novas aplicações, como as metodologias para avaliação genética e seleção com base no valor genômico (CAETANO, 2009).

Existem “SNP chips” com cerca de 4 milhões de SNPs destinados ao estudo em seres humanos, mais de 100 000 em plantas e em bovinos chips com mais de 700 mil marcadores(ILLUMINA INC., 2012), como exemplo podemos citar o projeto: *Estratégias genéticas para produção de carne bovina de qualidade*

(“Portal Embrapa”, [s.d.]) . O mesmo produziu 796 novilhos, descendentes de 34 touros selecionados para amostrar a diversidade da raça Nelore, os quais foram avaliados para eficiência alimentar e qualidade da carne, bem como submetidos à análise de associação genômica com um painel de SNPs (ILLUMINA INC., 2012).

2.3 Importância de alguns estudos genéticos

a) Estrutura da população

No ponto de vista genético uma população é definida como um grupo de indivíduos da mesma espécie, coexistindo e reproduzindo em uma área comum (ELER, 2017)**.** Tal categoria de população na genética é denominada população Mendeliana e é de fundamental importância nos estudos de genética de populações. Genética de populações, por outro lado, é um ramo da ciência com enfoque no estudo da distribuição dos genes nas populações, além de como as frequências gênicas e genotípicas são mantidas ou alteradas ao longo das gerações. Muitos estudos teóricos foram realizados a partir de informações de um, ou poucos locos, mas com a informação ampla de marcadores molecular podemos validar vários aspectos teóricos e situações práticas e aplicadas do melhoramento genético.

Denominamos frequência gênica, a proporção dos diferentes alelos de um gene na população. Já a frequência genotípica é o percentual com que cada um dos possíveis genótipos se encontra na população em estudo. Estas informações são a base para estudo da diferenciação e diversificação de espécies e, portanto, a geração de informações elaboradas sobre fenômenos genéticos depende, fundamentalmente, destas informações básicas e, portanto, merecem atenção especial quando se manipula grandes volumes de dados.

b) Análise genética em grandes conjuntos de dados

Dentre as contribuições da genética molecular para o melhoramento animal e vegetal, a utilização direta das informações de DNA no processo de identificação de indivíduos geneticamente superiores ganha destaque. A seleção genômica ampla (Genome Wide Selection-GWS), com a proposta de trabalhar com análise de um grande número de marcadores SNPs foi, portanto, idealizada. Em contrapartida, a utilização de tais informações não é simples uma vez que o número de marcadores é muito maior que o número de animais genotipados e os marcadores são altamente correlacionados. Desta forma metodologias que

contemplam esta adversidade tornam-se fundamentais (MEYER; GARCIA, 2002), mas certamente é necessário utilizar grandes volumes de dados que contenha informações relevantes para as análises que se destinam.

A vasta quantidade de informações em um banco de dados não garante que o mesmo tenha qualidade apropriada para fins de processamento. Dados incompletos, ruídos, contradições devem ser investigadas para que garanta, no banco de dados, a dimensionalidade efetiva para que métodos apropriados possam ser aplicados. Assim, estudos sobre dimensionalidade em grandes bancos de dados são também muito relevantes para as pesquisas atuais.

2.4 Descrição da estrutura genética de uma população

2.4.1 Frequências alélicas e genotípicas de uma população

No estudo de genética de populações contamos com duas informações fundamentais: a frequência alélica (ou frequência gênica) e a frequência genotípica que são descritores básicos da estrutura da população. Assim, considerando apenas o gene A/a, define-se uma população de tamanho N constituída por N_{11} indivíduos AA, N_{12} Aa e N_{22} aa, como relacionado na tabela abaixo.

Tabela 1 – Frequência genotípica em uma população

Genótipos	N° de indivíduos	Frequência
AA	N_{11}	$D = {N_{11}}/{N}$
Aa	N_{12}	$H = {N_{12}}/{N}$
aa	N_{22}	$R = {N_{22}}/{N}$
Total	N	1

Desta forma temos que:

$$f(A) = p = \frac{n_A}{n} = \frac{2N_{11} + N_{12}}{2N} = D + \frac{1}{2}H$$

$$f(a) = p = \frac{n_a}{n} = \frac{2N_{22} + N_{12}}{2N} = R + \frac{1}{2}H$$

Onde $p + q = 1$ e $n_A + n_a = n = 2N$

2.4.2 Equilíbrio de Hardy-Weinberg

O equilíbrio de Hardy-Weinberg (ou princípio de Hardy-Weinberg) é a

base da genética de populações (CRUZ, 2012). Admitindo uma população suficientemente grande, na ausência de seleção, migração e mutação, o equilíbrio é atingido após uma geração de acasalamento ao acaso. A relação genotípica torna-se igual ao quadrado da frequência alélica e permanece inalterada mesmo após sucessivas gerações de acasalamento ao acaso.

Tabela 2. - Frequência genotípica e alélica numa população antes e após acasalamento ao acaso (aaa).

População Inicial (P0)	População após o " aaa"(P1)
f(AA)=D	$\mathbf{f(AA) = \boldsymbol{D_1} = \boldsymbol{p^2}}$
f(Aa)=H	$\mathbf{f(Aa) = \boldsymbol{H_1} = \boldsymbol{2pq}}$
f(Aa)=R	$\mathbf{f(Aa) = \boldsymbol{R_1} = \boldsymbol{q^2}}$
f(A)=p	$\mathbf{f(A) = \boldsymbol{p_1} = \boldsymbol{p}}$
f(a)=q	$\mathbf{f(a) = \boldsymbol{q_1} = \boldsymbol{q}}$

Para avaliar se determinada população, ou mais precisamente cada loco independente, encontra-se ou não em equilíbrio de Hardy-Weinberg podemos utilizar alguns testes estatísticos, tais como o teste de qui-quadrado, teste de razão de verossimilhança e teste exato de Fisher. Assim, no teste de qui-quadrado, para verificar se um único loco, com s alelos, se encontra em equilíbrio, as seguintes classes genotípicas e suas quantidades observadas e esperadas devem ser consideradas:

Tabela 3. -Número observado e esperado de genótipos em uma população em equilíbrio de Hardy-Weinberg.

Genótipo	N° Observado ($\boldsymbol{O_{ij}}$)	N° Esperado ($\boldsymbol{E_{ij}}$)
$\boldsymbol{A_1A_1}$	N_{11}	p_1^2N
$\boldsymbol{A_1A_2}$	N_{12}	$2p_1p_2N$
$\boldsymbol{A_1A_3}$	N_{13}	$2p_1p_3N$
…	…	…
$\boldsymbol{A_{s-1}A_s}$	$N_{s-1,s}$	$2p_{s-1}p_sN$
$\boldsymbol{A_sA_s}$	$N_{s,s}$	p_s^2N
Total	N	N

Devemos verificar, portanto se o valor observado é estatisticamente igual ao esperado. As hipóteses testadas neste caso são:

H_0: A população se encontra em equilíbrio;

H_1: A população se encontra fora do equilíbrio;

A estatística qui-quadrado é dada por:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j \geq i}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Sendo,

O_{ij}: o número de indivíduos observados com genótipo $A_i A_j$.

E_{ij}: o número de indivíduos esperado com o genótipo $A_i A_j$.

Esta estatística está associada a $gl = n_c - r$ graus de liberdade, sendo:

n_c: o número de classes genotípicas avaliadas. Para s alelos, é dada por $n_c = \frac{S(S+1)}{2}$.

r: número de informações necessárias para determinar os valores esperados no teste de qui-quadrado. Para s alelos é necessário informação sobre o número total de indivíduos (N) e os valores das frequências de s-1 alelos, ou seja $r = 1 + (s - 1) = s$.

Desta forma, $gl = \frac{s(s-1)}{2}$.

Definidos o número de graus de liberdade (v) e o nível de significância (α) pode-se então obter o valor do qui-quadrado tabelado (χ_t^2). Para a decisão do teste compara-se o valor de qui-quadrado calculado (χ_c^2) ao valor tabelado (χ_t^2), adotando-se a seguinte regra:

Se $\chi_c^2 \leq \chi_t^2$, *então não rejeita-se* H_0;

Se $\chi_c^2 > \chi_t^2$, *então rejeita-se* H_0.

2.5 Análise da dimensionalidade por Componentes Principais

Proposta inicialmente por PEARSON (1901) a técnica dos componentes principais objetiva a simplificação estrutural de variação dos dados por meio da transformação de um conjunto original de variáveis em outro conjunto, os componentes principais, de dimensões equivalentes (HONGYU; JORGE; JUNIOR, 2015). Segundo Mingoti (2005), a técnica busca expressar a estrutura de variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio

da construção de combinações lineares, independentes entre si, das variáveis originais. Estas combinações são chamadas de componentes principais.

Muitas vezes nos deparamos com a necessidade de "redução do número de variáveis a serem avaliadas e interpretação das combinações lineares construídas", ou seja, expressar a informação contida nas p-variáveis originais em k (k<p) componentes principais não correlacionados. Portanto, o sistema de variabilidade do vetor aleatório que contém as k componentes principais representará a variabilidade contida nos dados inicias.

Seja $X = (X_1, X_2, \dots, X_p)'$ um vetor aleatório com vetor de médias $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ e matriz de covariâncias Σ_{pxp}. Sejam $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ os autovalores da matriz Σ_{pxp}, com seus respectivos autovetores normalizados $e_1, e_2, \dots, e_p$.

Definição 1: O j-ésimo componente principal da matriz $S,\ j = 1,2, \dots, p$, é definido como:

$$Y_j = e_j' X = e_{j1} X_1 + e_{j2} X_2 + \cdots + e_{jp} X_p.$$

Com esperança e variância dadas, respectivamente, por:

$$E[Y_j] = e_j' \mu = e_{j1}\mu_1 + e_{j2}\mu_2 + \cdots + e_{jp}\mu_p$$

$$Var[Y_j] = e_j' \Sigma_{pxp} e_j = \lambda_j$$

Temos que $Cov[Y_j, Y_k] = 0,\ j \neq k$. A variância de um componente principal Y_j é dada por λ_j. Os autovalores estão ordenados em ordem decrescente, desta forma o primeiro componente é o de maior variabilidade e o p-ésimo é o de menor.

Na prática, Σ_{pxp} é desconhecida e precisa ser estimada pela matriz de covariâncias amostral, S_{pxp}, calculada a partir dos dados amostrais como segue:

$$S_{pxp} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}.$$

Sendo $s_{ij} = s_{ji},\ j \neq i$, e s_{ii} a covariância amostral e s_{ij} variância amostral, definidas respectivamente por:

$$s_{ii} = \frac{\sum_{l=1}^{n} (X_{il} - \bar{X}_i)^2}{n-1}$$

$$s_{ij} = \frac{\sum_{l=1}^{n} (X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)}{n-1}.$$

Considere $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ os autovalores da matriz S_{pxp} e $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p$ seus respectivos autovetores normalizados. O j-ésimo componente principal estimado é dado por:

$$\hat{Y}_j = \hat{e}_j' X = \hat{e}_{j1} X_1 + \hat{e}_{j2} X_2 + \cdots + \hat{e}_{jp} X_p.$$

Os componentes principais são obtidos a partir da matriz de covariâncias e, por conta disso, sofrem influência das variáveis de maior variância, podendo ocorrer uma discrepância muito acentuada entre essas variâncias causada muitas vezes pela unidade de medida. Uma solução para evitar que este problema influencie nos resultados de agrupamento é a padronização dos dados. Em estatística, a padronização de uma variável refere-se a subtrair de cada valor a sua média e dividi-lo pelo seu desvio-padrão, e assim, os dados padronizados terão médias iguais a 0 e variâncias iguais a 1. Neste caso, os componentes principais são obtidos a partir da matriz de covariâncias da matriz R, sendo R a matriz das variáveis originais, ou seja, a matriz X, padronizadas.

Segundo a proposta de Jolliffe et al. 1973 descarta-se as variáveis com maiores coeficientes (em valor absoluto) com relação a componentes que possuam menores variâncias. Quando uma variável está relacionada ao maior coeficiente em mais de um componente recomenda-se que seja feito o descarte apenas desta variável, que não seja descartada outra varável do componente baseada em seus coeficientes. Nos estudos de diversidade genética descarta-se variáveis de componentes associados a autovalores nulos, ou segundo que possuir magnitude inferior a 0,7 (CRUZ; MEDEIROS, FÁBIO FERREIRA PESSONI, 2011).

3 Bibliografia

AZEVEDO, C. F. Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos. 2012.

CAETANO, A. R. Marcadores SNP: Conceitos básicos, aplicaões no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, n. SUPPL. 1, p. 64–71, 2009.

CAIXETA, E. T.; BORÉM, A. **Marcadores Moleculares**. Viçosa: UFV, 2006.

CRUZ, C. D. **Princípios de Genética Quantitativa**. Viçosa: UFV, 2012.

CRUZ, C. D.; MEDEIROS, FÁBIO FERREIRA PESSONI, L. A. **Biometria Aplicada ao estudo da diversidade genética**. Suprema ed. Visconde do Rio Branco: [s.n.].

ELER, J. P. **Teorias e métodos em melhoramento genético animal: Seleção**. Pirassununga: [s.n.].

HONGYU, K.; JORGE, G.; JUNIOR, D. O. Análise de Componentes Principais : resumo teórico , aplicação e interpretação. **Engineering and Science**, v. 1, p. 83–90, 2015.

ILLUMINA INC. BovineHD Genotyping BeadChip. **Pub. No. 370-2010-018**, v. WG-450-100, p. 1–4, 2012.

JOLLIFFE, A. I. T. et al. Discarding Variables in a Principal Component Analysis . II : Real Data Author (s): I . T . Jolliffe Published by : Blackwell Publishing for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2346300. v. 22, n. 1, p. 21–31, 1973.

Julia Micro-Benchmarks. , 2018. Disponível em: <https://julialang.org/benchmarks/>. Acesso em: 12 maio. 2018

LAURIE, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, v. 34, n. 6, p. 591–602, 2010.

MEYER, A. S.; GARCIA, A. A. F. **Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes.** Piracicaba: [s.n.].

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada**. Brlo Horizonte: UFMG, 2005.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

PINTOS, E. et al. **Modeling of RAFT Polymerization Processes Using an Efficient Monte Carlo Algorithm in Julia**. [s.l: s.n.]. v. 55

Portal Embrapa. Disponível em: <https://www.embrapa.br/pecuaria-sudeste/busca-de-projetos/-/projeto/207404/bases-moleculares-da-qualidade-da-carne-e-eficiencia-alimentar-na-raca-nelore>. Acesso em: 31 jan. 2018.

RESENDE, M. D. V. DE; SILVA, F. F. E; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa: Suprema, 2014.

RESENDE, M. D. V; SILVA, F. F.; RESENDE JÚNIOR, M. F. R. Seleção Genômica Ampla (GWS). **Biotecnologia aplicada ao melhoramento de plantas**, v. 1, p. 151–188, 2013.

CAPÍTULO 01

Golias: Análise estatística e biométrica de grandes conjuntos de dados de SNPs usando as linguagens Julia e R.

RESUMO

Oliveira, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Golias: Análise estatística e biométrica de grandes conjuntos de dados de SNP's utilizando os ambientes de programação Julia e R.** Orientador: Cosme Damião Cruz. Coorientador: Moysés Nascimento.

O Golias é um software livre destinado ao processamento de dados de alta dimensão, sendo compatível e integrado com os ambientes de programação R e Julia, ambos gratuitos. O Golias utiliza o poder de processamento da linguagem Julia e de pacotes do R otimizados em código C. Este software possibilita ao usuário realizar todas as análises sem ter qualquer conhecimento em programação. Em sua primeira versão, estão disponíveis procedimentos de controle de qualidade de marcadores moleculares SNPs como call rate, MAF, teste de equilíbrio de Hardy-Weinberg, descarte de variáveis via componentes principais; um procedimento para imputação de informações genotípicas perdidas e algumas técnicas de modelagem utilizando métodos bayesianos. O Golias está disponível em português; pode ser baixado da Internet (https://www.dropbox.com/golias) sendo compatível com sistema operacional Windows.

Palavras-chave: software, Big Data, análise de dados, biometria, GWS.

ABSTRACT

OLIVEIRA, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, February, 2018. **Golias: Statistical and biometric analysis of large SNP data sets using the Julia and R programming environments.** Adviser: Cosme Damião Cruz. Co-Adviser: Moysés Nascimento.

Golias is free software for high-quality data processing, being compatible and integrated with the R and Julia programming environments, both free. Golias uses the processing power of the Julia language and R-optimized C-packets. This software enables the user to perform all analyzes without any programming knowledge. In its first version, procedures are available for the quality control of molecular markers SNPs such as call rate, MAF, Hardy-Weinberg equilibrium test, variables discarding via main components; a procedure for imputation of lost genotypic information and some modeling techniques using Bayesian methods. Golias is available in Portuguese and can be downloaded from the Internet (https://www.dropbox.com/golias) being compatible with Windows operating system.

Key words: software, Big Data, data analysis, biometrics, GWS.

1 Introdução

O uso de informações moleculares tem gerado grandes contribuições no melhoramento genético. Informações no nível do DNA levem a um ganho genético mais rápido do que o obtido com base apenas em dados fenotípicos. (MEUWISSEN; HAYES; GODDARD, 2001).

Matrizes de alta densidade contendo milhares de informações genotípicas a cerca de indivíduos de determinada população torna-se cada vez mais frequentes nas diversas áreas da ciência. A tecnologia de SNP-Chip com a possibilidade de gerar rapidamente grande quantidade de informações ganha destaque diante do exposto. Porém deve-se atentar para a qualidade dos dados pois este fator poderá impactor nos resultados das analyses gerando resultados espúrios. Portanto adotar métodos de controle e garantia da qualidade torna-se imperativo para o sucesso da abordagem por meio de modelos de GWS (LAURIE et al., 2010) .

Apesar da ampla informação sobre o indivíduo estes conjuntos de dados trazem para dentro de procedimentos biométricos desafios referentes a dimensionalidade e multicolinearidade. Além disto, demandam grande eficiência de procedimentos computacionais pela necessidade de manipulação de matrizes de alta dimensão. Tais desafios têm sido confrontados por meio de métodos paramétricos de seleção genômica, destacando os Bayesianos (PÉREZ; DE LOS CAMPOS, 2014) e métodos fundamentados em inteligência computacional tais como redes neuras, árvores aleatória e máquinas de vetor suporte.

Para realizar o controle de qualidade e realizar os diversos processamentos destes dados, há disponível diferentes aplicativos gerados a partir de várias linguagens de programação tais como C++, R, Python, Java, Julia, MATLAB, dentre outros. Porém nem sempre é fácil decidir por qual optar, uma vez que para trabalhar com quantidades massivas de dados o desempenho do aplicativo na linguagem utilizada passa a ser altamente relevante pois influencia diretamente no tempo gasto para realizar o processo.

O R (https://www.r-project.org/), é um ambiente de programaçãoamplamente utilizado na área acadêmica. Conta com grande quantidade de pacotes já disponibilizados além de diversos modelos. De maneira geral, o R é uma linguagem destinada à computação estatística e geração gráficos em análises estatísticas. Ele dispõe de ferramentas estatísticas tais como modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento,

dentre outros. Para tarefas computacionais que exigem maior desempenho, destacam-se as linguagens C, C ++ e o código Fortran. Scripts desenvolvidos nesta linguagens podem ser vinculados ao R e chamados, em tempo de execução, permitindo que usuários possam escrever código, em linguagens externas, para manipular objetos diretamente no R ("R: The R Project for Statistical Computing", [s.d.]).

Outra linguagem, ou ambiente de programação, que tem ganhando destaque é a linguagem Julia (http://julialang.org/). Projetada para alcançar alto desempenho em computação numérica e científica, Julia é vista como uma alternativa de alto desempenho, fácil sintaxe e de código aberto para linguagens como MATLAB, R e Python. (PINTOS et al., 2016) ("The Julia Language", [s.d.]).

Tendo em vista a potencialidade dos recursos computacionais, este estudo tem por objetivo apresentar o software Golias. Desenvolvido com o objetivo de oferecer a pesquisadores que não possuem prática em programação uma interface simples associada ao grande potencial de processamento das linguagens Julia e R sem a necessidade de ter qualquer conhecimento em programação nas mesmas.

Ele utiliza códigos baseados em Julia para realizar o pré-processamento dos dados. Há disponível técnicas como Call Rate, MAF, Teste de equilíbrio de Hardy-Weinberg e descarte de variáveis via análise de componentes principais destinados ao controle de qualidade dos marcadores além de imputação de dados. Para a modelagem o Golias utiliza o pacote BGLR ('Bayesian Generalized Linear Regression') na predição de valores genéticos a partir de informações fenotípicas e genotípicas de uma população.

O BGLR (DE LOS CAMPOS; PEREZ-RODRIGUEZ, 2016) é um pacote que possui métodos de encolhimento bayesiano e variável para regressões de alta dimensão. Está disponível para a linguagem R e, mais recentemente, implementados pelos autores para linguagem Julia. A versão R do BGLR é altamente otimizada com o uso do código C compilado proporcionando ao pacote um desempenho, em termos de velocidade computacional, muito superior ao que poderia ser obtido usando apenas o código R.

2 Material e Métodos

O software Golias está disponível para sistemas operacionais Windows. Ele utiliza para realizar o processamento dos dados as linguagens Julia e R, desta forma é necessário que o usuário as tenha instalado em seu computador.

O software está disponível em português, todos os procedimentos implementados contam com um arquivo exemplo, o que tornará simples para o usuário preparar os arquivos de entrada. Juntamente com os arquivos de instalação está disponível um arquivo de instruções de instalação, a qual é feita de maneira muito simples.

Procediemntos disponíveis no Golias

O Golias possui, atualmente, procedimentos voltados para análise de dados de marcadores SNPs do tipo codominantes. Os procedimentos vão desde técnicas voltadas ao controle de qualidade dos marcadores a técnicas de modelagem utilizadas na predição de valores genéticos.

I. Descarte de variáveis via Análise de Componentes Principais (PCA)

Este procedimento pode ser utilizado para reduzir a dimensionalidade do conjunto de dados que o usuário tem disponível. A grande maioria dos dados oriundos de SNP's possuem um número de marcadores muito maior que o de indivíduos genotipados, desta forma a matriz de dados terá o número de colunas muito maior que o de linhas.

A técnica de descarte de variáveis utilizando PCA permite identificar os marcadores mais informativos (JOLLIFFE et al., 1973). Nos estudos de diversidade genética, descarta-se variáveis de componentes associados a autovalores nulos, ou que possuir magnitude inferior a 0,7(CRUZ; MEDEIROS, FÁBIO FERREIRA PESSONI, 2011). Desta forma este valor é fixado para este procedimento e utilizado como limite inferior para efetuar o descarte dos marcadores.

II. Controle de qualidade de SNP's

Há disponível no Golias duas metodologias para efetuar o controle de qualidade do banco de dados, *Call Rate* e frequência do alelo mais raro (MAF).

a) Call Rate

Call Rate (CR) refere-se à proporção de sucesso no retorno de genótipos. Desta forma esta medida é utilizada para eliminar marcadores que possuam um grande número de genótipos perdidos. Dado que **A** indivíduos apresentam genótipos ausentes e **P** retornam genótipos no processo de genotipagem temos que:

Estimador: $CR = 1 - \frac{A}{P}$

b) Frequência do alelo mais raro (MAF)

A MAF é uma medida que está ligada a variação alélica na população, uma vez que alelos pouco frequentes são pouco informativos, não apresentando assim relevância genética na população.

Considerando uma população bialélica de **N** indivíduos e alelos $\boldsymbol{A_1}$ e $\boldsymbol{A_2}$. O estimador para a MAF é dado por:

$$MAF = \min[p, q]$$

Sendo **p** a frequência do genótipo $\boldsymbol{A_1A_1}$ e **q** a frequência do genótipo $\boldsymbol{A_2A_2}$.

Geralmente opta-se por eliminar marcadores cuja **MAF** seja inferior a 1% ou 5%. Tendo em vista o call rate de cada marcador, geralmente, opta-se por eliminar marcadores cuja **CR** seja menor que 95% ou 90%.

Por default o Golias elimina marcadores cuja **MAF** seja inferior a 5% e para o call rate elimina marcadores cujo **CR** sejam inferiores a 95%. O usuário tem a opção de definir o nível de significância, caso não queira utilizar os pré-estabelecidos no software.

c) Equilíbrio de Hardy-Weinberg

O equilíbrio de Hardy-Weinberg (EWH) é a base da genética de populações (Cruz, 2012). È utilizado para verificar se as frequências genotípicas observadas estão de acordo com o esperado segundo o EWH.

2.1.1 Imputação de dados genotípicos

A imputação dos dados perdidos é feita utilizando o conceito de valor esperado. Considere **W** o vetor de informações de n indivíduos para um determinado marcador. W é então codificada da seguinte forma:

$$W_i = \begin{cases} 0, se\ o\ indivíduo\ apresenta\ o\ genótipo\ A_1A_1 \\ 1, se\ o\ indivíduo\ apresenta\ o\ genótipo\ A_1A_2 \\ 2, se\ o\ indivíduo\ apresenta\ o\ genótipo\ A_2A_2 \end{cases}$$

Considerando **p** a frequências do alelo A_2 temos então que,

$$E(W) = \sum_{i=1}^{3} w_i P(w_i) = 0q^2 + 2pq + 2p^2 = = 2p(1-p) + 2p^2$$

$$= 2p - 2p^2 + 2p^2 = 2p$$

A variável Indicadora W possui distribuição Binomial e que seu valor esperado é dado por $E(W) = 2p$. Pode-se então utilizar este estimador para imputar valores

perdidos do marcador nos indivíduos. Para tanto segue-se a seguinte regra (Resende, Silva, and Azevedo 2014):

$$W_i = \begin{cases} Se\ 2p \leq 0{,}5\ então\ W_i = 0 \\ Se\ 0{,}5 < 2p \leq 1{,}5\ então\ W_i = 1 \\ Se\ 1{,}5 < 2p \leq 2\ então\ W_i = 2 \end{cases}$$

III. Modelos utilizados para Seleção Genômica

Como dito anteriormente nos estudos de genômica, torna-se cada vez mais comum conjuntos de dados de alta densidade tal que o número de marcadores é muito maior que o número de indivíduos genotipados, o que gera vários desafios estatísticos e computacionais para implementação de regressões. Alguns destes desafios podem ser confrontados utilizando métodos bayesianos (PÉREZ; DE LOS CAMPOS, 2014).

Para proceder com o ajuste dos modelos o Golias utiliza o pacote BGLR(DE LOS CAMPOS; PEREZ-RODRIGUEZ, 2016) desenvolvido para o software R. Este pacote, muito conhecido por usuários do software R, implementa uma grande coleção de métodos bayesianos. Está disponível no software os seguintes modelos de regressão: Bayesian Ridge Regression, BayesA, BayesB.

Utilizando o Golias

Ao instalar o software pela primeira vez em seu computador, é de extrema importância que seja feita a instalação das linguagens Julia e R e dos pacotes que são utilizados nos procedimentos. Desta forma no menu Integração, está disponível as opções “Instalar Pacotes Julia” e “Instalar Pacotes R”.

O software possui um arquivo exemplo, para todos os procedimentos implementados. Com isto o usuário poderá observar a forma com que os dados estão dispostos e organizados.

Além disto ao definir o arquivo a ser processado, pressionando o botão “Carregar” é exibido na janela do software as seguintes informações: número de linhas, número de colunas e é listado os elementos distintos do banco de dados. É exibido também as 4 primeiras colunas e 5 primeiras linhas do conjuto de dados. Esta aplicação é útil para que o usuário certifique que seus dados foram lidos de forma correta e também verificar se existem erros em seus dados tal como algum erro de digitação.

Após concluir o procedimento, na janela de texto é exibido um breve resumo

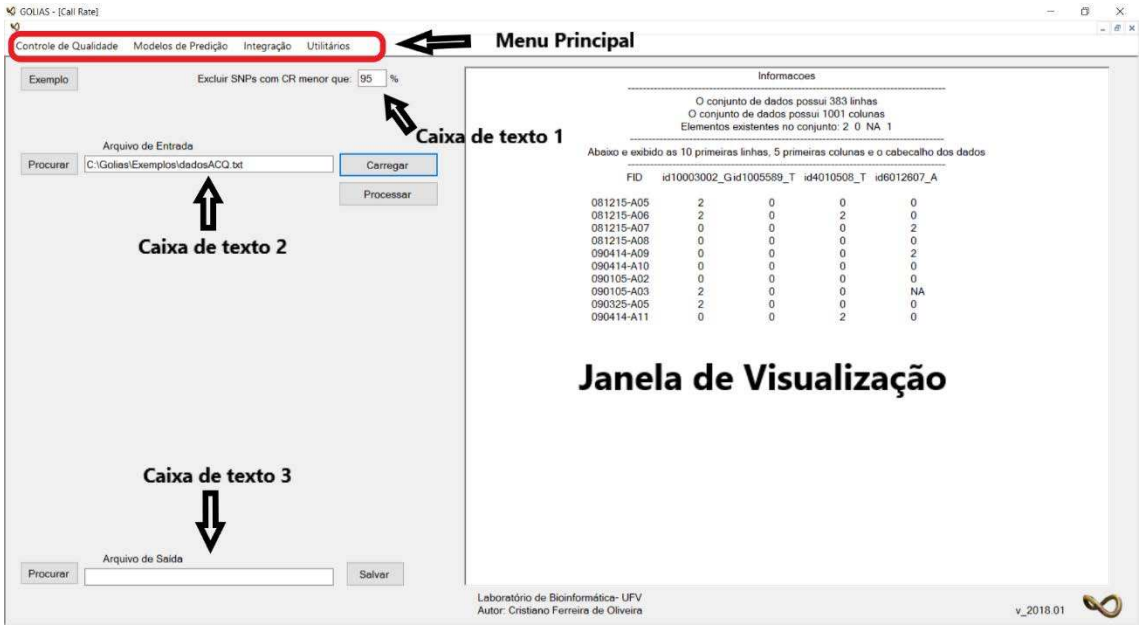
ou os resultados do que foi feito. Para salvar arquivos processados, basta definir o diretório para salvá-lo e clicar no botão salvar. Os arquivos de saída sempre são salvos como uma extensão do tipo txt.

Exemplo de aplicação:

Neste exemplo será feita a prática de controle de qualidade de um conjunto de marcadores. O conjunto de dados utilizados está disponível no software como um arquivo exemplo. Serão realizados os procedimentos Call Rate e MAF.

2.1.2 Call Rate

Inicialmente considera-se a tela inicial de acesso ao aplicativo conforme ilustrado a seguir:



i. No menu principal clique em *"Controle de qualidade"* , selecione a opção *"Call Rate"*.

ii. Na aba que será aberta clique no botão "Exemplo". Observe as opções disponíveis. Na caixa de texto 2 *"Arquivo de Entrada"* estará preenchida com o caminho do arquivo exemplo.

Para selecionar um arquivo pode-se simplesmente preencher a caixa de texto com a informação do diretório e o nome do arquivo ou então clicando no botão *"Procurar"* situado na parte superior e na janela que irá abrir procurar a pasta do arquivo e selecioná-lo.

Por default o Golias elimina marcadores com Call Rate (CR) inferior a 95%, o valor é especificado na caixa de texto 1. Para alterar este valor basta clicar na

caixa de texto1 e entrar com o valor desejado.

iii. Clique no botão *"Carregar"*. Será exibido na *"Janela de Visualização"* umresumo do conjunto de dados, contendo informações a respeito do número de linhas, número de colunas, elementos presentes no arquivo de dados e as 5 primeiras linhas, 4 primeiras colunas e o cabeçalho dos dados (O usuário pode verificar se o seu conjunto foi lido corretamente. Além disso com o arquivo exemplo ele poderá ver como o arquivo deve estar ordenado).

iv. Clique no botão "Processar". Ao término do processo será exibido na "Janela de

Visualização" as informações referentes ao número de linhas e colunas do conjunto de dados inicial, número de linhas e colunas do conjunto de dados final e valor do CR adotado para eliminar os marcadores;

v. O próximo passo consiste em salvar o arquivo. Para isto basta clicar no botão *"Procurar"* na parte inferior da tela. Na janela que abrirá vá até a pasta na qual deseja salvar o arquivo processado e clique no botão "OK". A caixa de texto 3 *"Arquivo de Saída"* será então preenchida com o diretório onde o arquivo

será salvo e também com o nome do arquivo que por default será *dadosCR.txt.* Para alterar o nome basta trocar o nome *dadosCR* pelo nome desejado.

Clique no botão *"Salvar"*. Será exibida na tela uma mensagem informando que o arquivo foi copiado com sucesso. Clique em ok. O arquivo processado se encontra no diretório definido para salvá-lo.

2.1.3 MAF

i. No menu principal clique em controle de qualidade, selecione a opção *"MAF"*.

ii. Na aba que será aberta clique no botão "*Procurar*" situado na parte superior e selecione o arquivo salvo do procedimento anterior.

Por default o Golias elimina marcadores com MAF inferior a 5%, o valor é especificado na caixa de texto 1. Para alterar este valor basta clicar na caixa de texto1 e entrar com o valor desejado.

Caso queira ver o resumo do conjunto de dados clique no botão *"Carregar"*.

iii. Clique no botão *"Processar"*. Ao término do processo será exibido na *"Janela de Visualização"* as informações referentes ao número de linhas e colunas do conjunto de dados inicial, número de linhas e colunas do conjunto de dados final e valor do MAF adotado para eliminar os marcadores;

iv. Salve o arquivo processado.

2.1.4 Imputação dos dados Genotípicos

i. No menu principal clique em controle de qualidade, selecione a opção *"Imputar Dados"*.

ii. Na aba que será aberta clique no botão "*Procurar*" situado na parte superior e selecione o arquivo salvo do procedimento anterior. Clique no botão *"Carregar"*. Observe no resumo da *"Janela de Visualização"* que os genótipos estão codificados com os valores 0 (genótipo A_1A_1) ,1 (genótipo A_2A_2) e 2 (genótipo A_1A_2). Além disto existe a presença de dados perdidos os quais são codificados como NA (ausência da informação genotípica).

iii. Clique no botão *"Processar"*. Ao término do processo será exibido na *"Janela de Visualização"* a mensagem "Dados imputados com sucesso";

iv. Salve o arquivo com os dados imputados.

3 Resultados e Discussão

Golias é um software destinado a análise de grandes conjuntos de marcadores. A linguagem Julia, utilizada para realizar o pré-processamento dos dados e o BGLR-R para as análises genômicas, proporciona a ele grande poder de processamento

Possui procedimentos utilizados para garantir a qualidade dos dados, descarte de variáveis e também modelagem para predição de valores genéticos utilizando métodos bayesianos. Gratuito e fácil de usar, não demanda do usuário qualquer conhecimento em programação podendo ser utilizado tanto em instituições públicas como privadas.

4 Agradecimentos

O autor agradece a CAPES pelo apoio financeiro.

5 Referências Bibliográficas

CRUZ, C. D.; MEDEIROS, FÁBIO FERREIRA PESSONI, L. A. **Biometria Aplicada ao estudo da diversidade genética**. Suprema ed. Visconde do Rio Branco: [s.n.].

DE LOS CAMPOS, G.; PEREZ-RODRIGUEZ, P. BGLR: Bayesian generalized linear regression. R package version 1.0.5. p. https://CRAN.R-project.org/package=BGLR, 2016.

JOLLIFFE, A. I. T. et al. Discarding Variables in a Principal Component Analysis . II : Real Data Author (s): I . T . Jolliffe Published by : Blackwell Publishing for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2346300. v. 22, n. 1, p. 21–31, 1973.

LAURIE, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, v. 34, n. 6, p. 591–602, 2010.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

PINTOS, E. et al. **Modeling of RAFT Polymerization Processes Using an Efficient Monte Carlo Algorithm in Julia**. [s.l: s.n.]. v. 55

R: The R Project for Statistical Computing. Disponível em: <https://www.r-project.org/>. Acesso em: 25 jan. 2018.

The Julia Language. Disponível em: <https://julialang.org/>. Acesso em: 2 dez. 2017.

CAPÍTULO 02

Eficiência no processamento de dados por meio das linguagens R e Julia no controle de qualidade e predição de valor genético em conjuntos de dados de SNP’ s de alta densidade.

RESUMO

Oliveira, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Eficiência no Processamento de dados por meio das linguagens R e Julia no controle de qualidade e predição de valor genético em conjuntos de dados de SNP' s de alta densidade.** Orientador: Cosme Damião Cruz. Coorientador: Moysés Nascimento.

Em programas de melhoramento genético, é crescente o uso de milhares de informações moleculares dispostos em matrizes de alta densidade, que necessitam de controle de qualidade e uso de metodologias adequadas paras realizar a predição dos valores genéticos. Diante do exposto, o objetivo deste trabalho foi verificar o desempenho das linguagens Julia e R para o controle de qualidade de matrizes de SNPs de alta dimensão além de comparar o desempenho do pacote BGLR (*Bayesian Generalized Linear Regression*) nos ambientes de programação Julia e R. Os resultados mostraram que a linguagem Julia e R tiveram desempenhos similares no processo de controle de qualidade dos dados e que o BGLR-R foi mais eficiente tanto em termos de tempo computacional quanto em acurácia seletiva e preditiva.

Palavras-chave: BGLR, programação, análise de dados, biometria, GWS.

ABSTRACT

OLIVEIRA, Cristiano Ferreira, M.Sc., Universidade Federal de Viçosa, February, 2018. **Efficiency in Data Processing through R and Julia languages in quality control and prediction of genetic value in data sets of high density SNP's.** Adviser: Cosme Damião Cruz. Co-Adviser: Moysés Nascimento.

In genetic breeding programs, the use of thousands of molecular information arranged in high density matrices, which require quality control and the use of appropriate methodologies to predict genetic values, is increasing. The objective of this work was to verify the performance of the Julia and R languages for the quality control of matrices of high SNPs in addition to comparing the performance of the BGLR (Bayesian Generalized Linear Regression) package in the Julia and R programming environments. The results showed that the Julia language and R had similar performances in the process of data quality control and that BGLR-R was more efficient both in terms of computational time and accuracy.

Key words: BGLR, programming, data analysis, biometrics, GWS.

1 Introdução

A potencialidade do uso de marcadores no melhoramento genético levou ao estabelecimento de métodos eficazes de detecção e quantificação de polimorfismo em nível de DNA tal como SSCP-SNP (Marcador molecular co-dominante e robusto baseado em SNPs) que permitem caracterizar grande número de locos (CAIXETA; BORÉM, 2006). Tais métodos constituem avanços tecnológicos produzidos como resultado da evolução nas técnicas de determinação de marcas moleculares, deenominadas genotipagem de alta densidade, que geram, atualmente, grande volume de dados que, ao ser processados e analisados, contribui para tomadas de decisão e representam grande informação sobre o indivíduo a partir de seu DNA.

Por meio de informações fenotípicas agregadas às informações moleculares é possível predizer valores genéticos de forma mais acurada e, neste contexto, a abordagem por meio de modelos de seleção genômica ampla (GWS) se destaca, pois em seus modelos contemplam as informações de grandes números de marcadores amplamente distribuídos ao longo do genoma.

Conjuntos de dados de SNP' s de alta densidade são cada vez mais comuns nas diversas áreas da ciência. Como exemplo temos o conjunto de genótipos de gado bovino abrangente do genoma que possui mais de 777.000 SNP' s (ILLUMINA INC., 2012). A capacidade de gerar rapidamente grande quantidade de informações faz com que a tecnologia de SNP-Chip se destaque. Por outro lado a qualidade dos dados pode afetar os resultados das análises, gerando resultados espúrios, desta forma métodos de controle e garantia da qualidade se tornam fundamentais(LAURIE et al., 2010) .

Estes conjuntos possuem um número de marcadores muito maior que o número de indivíduos genotipados e, assim, apesar de proporcionar ampla informação sobre o indivíduo trazem para dentro de procedimentos biométricos questões importantes referentes a dimensionalidade e multicolinearidade. Adicionalmente, demandam grande eficiência de procedimentos computacionais pela necessidade de manipulação de matrizes de alta dimensão.

No contexto estatístico, soluções para problemas de alta dimensão tem sido abordado por métodos paramétricos de seleção genômica, destacando os

Bayesianos (PÉREZ; DE LOS CAMPOS, 2014) e métodos fundamentados em inteligência computacional tais como redes neuras, árvores aleatória e máquinas de vetor suporte (RESENDE; SILVA; AZEVEDO, 2014).

A escolha do recurso computacional impacta diretamente no tempo de processamento dos dados, portanto, tal assunto merece considerável atenção por parte do pesquisador. Para realizar o controle de qualidade e os diversos processamentos destes dados, tem sido disponibilizado diferentes aplicativos gerados a partir de várias linguagens de programação tais como C++, R, Python, Java, Julia, MATLAB, dentre outros. Porém nem sempre é fácil decidir por qual optar, uma vez que para trabalhar com quantidades massivas de dados o desempenho do aplicativo na linguagem utilizada passa a ser altamente relevante pois influencia diretamente no tempo gasto para realizar o processo.

Um ambiente de programação que ganha destaque na comunidade acadêmica é o R (https://www.r-project.org/), devido a quantidade de pacotes já disponibilizados e a diversidade de modelos que esta linguagem (ou ambiente de programação) oferece. De maneira geral, o R é uma linguagem destinada à computação estatística e geração gráficos em análises estatísticas. Ele dispõe de ferramentas estatísticas tais como modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, dentre outros. Para tarefas computacionais que exigem maior desempenho, destacam-se as linguagens C, C ++ e o código Fortran. Scripts desenvolvidos nesta linguagens podem ser vinculados ao R e chamados, em tempo de execução, permitindo que usuários possam escrever código, em linguagens externas, para manipular objetos diretamente no R (TEAM, 2013).

Outra linguagem que também vem ganhando destaque é a linguagem Julia (http://julialang.org/). Projetada para alcançar alto desempenho em computação numérica e científica, Julia é vista como uma alternativa de alto desempenho, fácil sintaxe e de código aberto para linguagens como MATLAB, R e Python. (PINTOS et al., 2016).

Neste trabalho foi comparada a eficácia dos ambientes de programação R e Julia para o pré-processmento de dadose do pacote BGLR-R e BGLR-Julia na predição de valores genéticos, para fins de melhoramento de planta tendo como referência a cultura do arroz com um banco de dados moleculares.

Para fins de comparação, a predição foi feita por meio do pacote

BGLR(‘Bayesian Generalized Linear Regression’) (PÉREZ; DE LOS CAMPOS, 2014). A escolha do arroz como espécie utilizada para o estudo é interessante, pois trata-se da principal fonte de alimento para mais de 3,5 bilhões de pessoas em todo o mundo, cerca de metade da população mundial. O arroz é o alimento básico na Ásia, onde 600 milhões de pessoas vivem em extrema pobreza. Além disto, sua importância está aumentando em regiões afetadas pela pobreza na África e América Latina (“IRRI - Research”, 2018). Estudos sobre a base genética da variação fisiológica, evolutiva e morfológica no arroz asiático (Oryza sativa) é imprescindível para o aumento da qualidade e sustentabilidade do suprimento de alimentos no mundo.

O pacote BGLR (PÉREZ; DE LOS CAMPOS, 2014) implementa métodos Bayesianos para regressões. Está disponível para a linguagem R e, mais recentemente, implementados pelos autores para linguagem Julia (https://github.com/gdlc/BGLR.jl). A versão R do BGLR (BGLR-R) é altamente otimizada com o uso do código C compilado proporcionando ao pacote um desempenho, em termos de velocidade computacional, muito superior ao que poderia ser obtido usando apenas o código R. Sua versão para linguagem Julia (BGLR-Julia) foi feita com o objetivo de alcançar a comunidade de usuários desta linguagem e visando alcançar um desempenho similar ou melhor do que o alcançado com BGLR-R (PEREZ-RODRIGUEZ; DE LOS CAMPOS, 2018)

Tendo em vista a potencialidade destes recursos computacionais, comparou-se o desempenho do pacote BGLR, nos ambientes R e Julia, na predição de valores genéticos a partir de informações fenotípicas e genotípicas de uma população de arroz, tomando como critério valores de acurácia de predição e tempo de processamento. Além disto, as linguagens foram utilizadas para o pré-processamento dos dados através do uso rotinas que permitam realizar o controle de qualidade dos marcadores por meio de técnicas como Call Rate e MAF, e então comparado o desempenho de cada uma em termos de velocidade de processamento.

2 Material e Métodos

2.1 Material

A comunidade científica vem disponibilizando vários bancos de informações, que constituem recursos genômicos, tais como sequências de genoma de alta qualidade, mapas densos de SNP’s e coleções de germoplasma com acesso públicos

de informações genômicas (ZHAO et al., 2011). Neste trabalho foram utilizadas as informações de um banco de dados de SNP's gerados a partir da experimentação com a cultura do arroz (*Oryza sativa) e* oriundos do trabalho relatado por Zhao *et al.*, 2011. Inicialmente foi acessado um banco de dados que continha informações genotípicas, relativas a 30122 marcadores, obtidas em uma população composta por 383 indivíduos. Considerando os alelos A_1 e A_2 as informações genotípicas foram codificados como 0, 1 ou 2 sendo,

$$\begin{cases} 0, & presença\ do\ genótipo\ A_1A_1 \\ 1, & presença\ do\ genótipo\ A_1A_2 \\ 2, & presença\ do\ genótipo\ A_2A_2 \end{cases}$$

(1)

Para fins comparativos foi selecionada para as análises biométricas a característica período de floração que corresponde ao tempo, em dias, da semeadura plantio até o ponto em que 50% das panículas estão floridas. Esta característica tem controle gênico quantitativo e está relacionada ao desenvolvimento da planta. Características. como o período para o florescimento tem sua expressão fenotípica influenciada por fatores ambientais como o foto período (PLANT et al., 2001). Desta forma a variação genotípica, fenotípica e ambiental se manifestam em diferentes magnitudes gerando, assim, grandes desafios para a seleção genômica (ZHAO et al., 2011).

2.2 Metodologias

a) Filtragem no banco de dados moleculares

Diante dos desafios de se ter valores acurados a partir de análises que envolvam mensurações fenotípicas e genotípicas, deve se ter a preocupação de se gerar informações fenotípicas confiáveis e dados genotípicos de qualidade. Assim, deu-se ênfase em procedimentos que permitam garantir a qualidade dos dados moleculares entendendo que é um passo crucial para o sucesso no processo de seleção genômica, uma vez que utilizar marcadores de baixa qualidade poderá comprometer a qualidade dos resultados.

Assim foram desenvolvidos scripts, em programação fundamentados nas linguagens Julia e R focando os critérios comumente utilizados para filtrar a baixa qualidade dos marcadores tais como valores de call rate e de MAF.

Call Rate do SNP:

Call rate (CR) ou taxa de atendimento na genotipagem de N indivíduos é dada por:

$$CR = 1 - \frac{A}{P}$$

Em que A representa o número de indivíduos que apresentam genótipo ausente (missing) e P representa o número de indivíduos que retornam genótipos no processo de genotipagem (not missing). Este valor é utilizado no controle de qualidade para eliminar marcadores que possuem um número significativo (CR inferior ao limite estabelecido pelo pesquisador) de informações genotípicas perdidas.

Classicamente, os marcadores com um call rate inferior a 95% são removidos de estudos adicionais, embora em alguns estudos tem-se estabelecido um limiar entre 95% e 99% (ANDERSON et al., 2010). Em nosso estudo foram eliminados indivíduos com call rate inferior a 95%.

MAF - Menor frequência alélica

A "Minor Allele Frequency" (MAF) refere-se à frequência à qual o alelo menos comum ocorre numa dada população. Geralmente, marcadores com MAF entre 1% e 5% na população são excluídos das análises posteriores, a fim de se aumentar o poder de detecção da associação genótipo-fenótipo (ANDERSON et al., 2010). Este parâmetro é utilizado para diferenciar polimorfismos comuns de variantes raras e erros de genotipagem. Para este processo de controle de qualidade estabelecemos um limiar de 5%.

b) Predição de valores genéticos

O uso de mapas de marcadores que cobre todos os cromossomos, torna possível a predição do valor genético mais acurada. Métodos que assumem uma distribuição prévia para a variância associada a cada segmento cromossômico forneceram previsões mais precisas de valores genéticos (MEUWISSEN; HAYES; GODDARD, 2001).

Desta forma, para avaliar a eficiência e desempenho dos ambientes de programação R e Julia na predição de valores genotípicos, foram utilizados os modelos de predição Bayes A e Bayes B disponíveis no pacote BGLR, utilizando os dados genotípicos após realizada a filtragem por meio de critérios de controle de qualidade.

A predição de valores genéticos genômicos, utilizando informações fenotípicas e genotípicas para cada indivíduo, foi estimada usando o modelo (MEUWISSEN; HAYES; GODDARD, 2001):

$$y = \mu 1_n + Xg + e$$

onde y é o vetor de valores fenotípicos; 1_n é um vetor de mesma dimensão de y com todos os elementos iguais a 1; μ é a média da característica de interesse; g é o vetor de efeitos genéticos aditivos dos marcadores; X é a matriz de incidência que relaciona os efeitos aditivos dos marcadores aos fenótipos; e é o vetor de erro aleatório do modelo tal que $e \sim N(0, \sigma_e^2)$ e σ_e^2 é a variância do erro.

A característica cujo valor genotípico foi predito refere-se ao período de floração de plantas de uma população de arroz. Também para os dados fenotípicos verificou-se que havia informações perdidas para alguns indivíduos, os quais foram excluídos do conjunto de dados fenotípico e genotípico restando 348 indivíduos para a predição.

Para comparar a eficiência do modelo de predição realizou-se procedimento de validação cruzada como ilustrado na Figura 1.

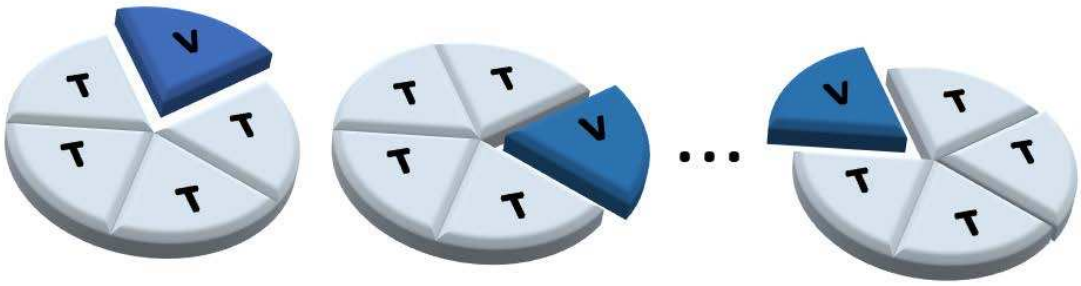


Figura 1: Esquema ilustrando o processo de validação cruzada, no qual o conjunto de dados foi dividido em 80% para treinamento (T) e 20% para validação (V).

Neste procedimento o conjunto de dados foi dividido de maneira sistemática, em 5 partes, onde cada uma continha aproximadamente 20% dos indivíduos e feito o procedimento de validação cruzada como descrito na figura 1.

A fim de verificar a relação entre o tempo de processamento e o número de marcadores foram utilizados conjuntos de dados de tamanhos distintos e feito um processo de validação cruzada para cada um destes. Para formar tais conjuntos, a matriz de informações genotípicas pós controle de qualidade (M) foi utilizada. O

primeiro era composto por todos os marcadores disponíveis, 26205 SNPs ou seja, a própria matriz M. Para formar o segundo, foram removidos os 5000 últimos marcadores de M, para o terceiro os 10000 últimos e assim sucessivamente. Por fim, foram formados seis conjuntos de tamanhos distintos contendo 26205, 21205, 16205, 11205, 6205 e 1205 SNPs.

c) Pacote BGLR disponíveis nos ambientes de programação R e Júlia

Estudos genéticos, nos quais, os fenótipos são explicados por um grande número de genes se destaca como um problema de aprendizagem estatística envolvendo dados de alta dimensão (PÉREZ; DE LOS CAMPOS, 2014). A implementação de regressões onde o número de parâmetros a serem estimados é muito superior ao número de observações apresenta vários desafios estatísticos e computacionais. A fim de confrontar tais desafios foi então desenvolvido o pacote BGLR para o R (BGLR-R) cujo algoritmo é baseado em um amostrador de Gibbs. Atualmente o pacote está disponível também para a Julia (BGLR-Julia).

Todos os argumentos da função BGLR possuem valores padrão, exceto o vetor de fenótipos. Por padrão um total de 1.500 ciclos de um amostrador de Gibbs (o valor padrão para o número de iterações) são executados, e as primeiras 500 amostras são descartadas (valor padrão para burn-in). Das amostras coletadas algumas são salvas no disco rígido em arquivos com extensão $*$.dat enquanto as mais recentes permanecendo na memória. Os meios de execução necessários para calcular estimativas das médias posteriores e dos desvios padrão posteriores são então atualizados. É utilizado por padrão um afinamento de 5 podendo ser modificado pelo usuário por meio do argumento thin. Ao fim do processo é retornado uma lista com estimativas e os argumentos usados na chamada.

Para o ajuste dos modelos utilizando o pacote foi fixado os argumentos, número de iterações, burn-in (número de amostras descartadas) e thin, respectivamente iguais a 100000, 10000 e 10.

A convergência do modelo foi verificada utilizando-se o critério de Heidel (HEIDELBERGER; WELCH, 1983), disponível no pacote do R, Coda (MARTYN et al., 2016).

d) Medidas comparativas de desempenho

Foram empregadas três medidas comparativas, acurácia seletiva, acurácia preditiva e tempo de processamento.

Acurácia seletiva: Esta medida pode ser obtida tanto para o conjunto de

informação submetida a treinamento quanto para a validação. Ela é dada pela correlação entra o valor fenotípico observado e o valor fenotípico predito pelo modelo,

$$r_{Y,\hat{Y}} = cor(Y, \hat{Y});$$

sendo Y os valores fenotípicos observados, $\hat{Y}$ os valores fenotípicos preditos por meio da regressão, e **r** o coeficiente de correlação de Pearson. O coeficiente de correlação de Pearson mede a relação linear entre as variáveis estudadas. Ele varia de -1 a 1, sendo que valores próximos de 1 indica forte relação linear positiva, próximos de 0, ausência de relação linear, e próximos de -1 forte relação linear e negativa.

Desta forma para correlação próxima de 1 indica que os maiores valores genéticos preditos estão associados aos maiores valores genéticos observados. Proporcionando assim, maior segurança para efetuar a seleção dos melhores indivíduos para a característica desejada utilizando o modelo ajustado.

Acurácia preditiva: É dada pelo erro quadrático médio estabelecido a partir das diferenças entre os valores fenotípicos observados e preditos pelo modelo,

$$EQM = \sum_{=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}$$

em que:

EQM é erro quadrático médio, y_i é valor observado para o i-ésimo indivíduo, $\hat{y}_i$ o valor predito para o i-ésimo indivíduo e ***n*** o número de indivíduos.

O erro quadrático médio é uma medida de dispersão, assim baixos valores indicam que o valor predito se encontra próximo do valor observado. Neste cenário isto acarretará em maior confiança em termos de predição e seleção usando o modelo, uma vez que os valores estão sendo preditos com precisão.

Tempo de processamento: O tempo de processamento é dado pelo tempo gasto para realizar o ajuste de um modelo. Para tanto foram utilizadas funções dentro de cada ambiente que permitiam medir o tempo gasto para executar uma rotina. No R utilizou-se a função “Sys.time()”, na linguagem Julia utilizou-se a função “tic() toq()”.

e) Recursos computacionais

Para realizar as análises foi utilizado um computador portátil com sistema operacional Microsoft Windows 10 Home, processador central Intel(R) Core(TM)

i7, adaptador de Vídeo NVIDIA GeForce 930M com memória de vídeo de 4.0 GB e 8 GB de memória RAM.

3 Resultado e Discussão

3.1 Filtragem no banco de dados moleculares;

Na seleção genômica o controle de qualidade dos marcadores é crucial para otimizar a acurácia dos resultados nos estudos (LAURIE et al., 2010)(TARANTO et al., 2016). Para o melhorista é funtamental dispor de ferramentas capazes de trabalhar com matrizes de alta dimensão, capazes de incorporar algoritmos matemáticos e estatíscos por meio de uma sintaxe simples e que permita submeter o conjunto de dados a uma avaliação rápida e eficiente de sua qualidade (FRANKE et al., 2016). Desta forma, foram utilizados scripts associados aos ambientes de programação Julia e R para realizar o controle de qualidade da matriz contendo as informações moleculares.

Assim, foi possível processar o conjunto de dados inicial que possuia informações de 30122 marcadores, e verificar o desempenho dos dois ambientes de programação para o mesmo. Após o processo de filtragem, utilizando um call rate de 95% e um MAF de 5%, este conjunto de dados foi reduzido considerávelmente, passando a ser composto por um total de 26205 marcadores utilizados nos procedimentos subsequentes.

Ambas as linguagens se mostraram eficientes para esta etapa, realisando o processo de filtragem em tempos computacionais muito próximos.

3.2 Comparação de desempenho do pacote BLGR nos ambientes R e Júlia

A convergência dos modelos foi testada utilizando-se o critério de Heidel. Verificou-se que os modelos convergiram em todos os cenários avaliados.

a. Acurácia seletiva

No melhoramento genético uma das principais atividades é a identificação de genótipos superiores para fins de seleção e, neste caso, duas medidas passam a ser de grande importância que são a acurácia seletiva e a acurácia preditiva (CAVALCANTI et al., 2012) (PEIXOTO; BHERING, 2016). A acurácia, de maneira geral, tem sido entendida como a medida da proximidade do valor predito e o valor real . Entretando, para fins de melhoramento, pode-se optar por medida de

acurácia seletiva que mede, não a proximidade de valores per se, mas o quanto que a discriminação de dois ou mais genótipos, com base no valor fenotípico, se aproxima da discriminação real destes genótipos (MORAIS JÚNIOR et al., 2013). Esta acurácia seletiva tem sido medida por meio de estatísticas de associação tais como covariâncias e correlações. Nas Tabelas 1 e 2 é apresentado os valores de acurácia seletiva, medida por meio do coeficiente de correlação de Pearson.

Tabela 1. Acurácia seletiva (coeficiente de correlação de Pearson) obtida em análise de predição de valores genéticos de plantas de arroz, utilizando o modelo *bayesA*, considerando diferentes conjuntos de informações moleculares.

Número de Marcadores	R		Julia	
	Treinamento	Validação	Treinamento	Validação
26205	0,95	0,61	0,83	0,48
21205	0,94	0,61	0,85	0,43
16205	0,93	0,60	0,81	0,40
11205	0,89	0,55	0,77	0,42
6205	0,88	0,53	0,77	0,37
1205	0,80	0,50	0,70	0,38

Tabela 2. Acurácia seletiva (coeficiente de correlação de Pearson) obtida em análise de predição de valores genéticos de plantas de arroz, utilizando o modelo *bayesB*, considerando diferentes conjuntos de informações moleculares

Número de Marcadores	R		Julia	
	Treinamento	Validação	Treinamento	Validação
26205	0,95	0,61	0,84	0,45
21205	0,94	0,61	0,88	0,42
16205	0,93	0,60	0,84	0,46
11205	0,90	0,56	0,79	0,36
6205	0,89	0,53	0,76	0,41
1205	0,82	0,51	0,71	0,39

Quando o objetivo é obter valores acurados em modelos de predição, deve se ter algumas preocupações básicas sendo o primeiro deles a obtenção de valores

confiáveis das variáveis respostas e das variáveis preditivas. Para a variável resposta devemos lembrar que o fenótipo é resultado da expressão genótipica e também do ambiente. Desta forma, características, como tempo de floração de plantas de arroz que é altamente influenciada pelo ambiente aumentam ainda mais o desafio que o melhorista terá de enfrentar, pois isto terá grande influência na acurácia do modelo ajustado (HUANG et al., 2012). Neste caso, a obtenção de valores fenotípicos deve ser feita utilizando delineamentos apropriados e realizando mensurações a partir da condução dos experimentos minimizando a influência de fatores externos.

Com releção às variáveis explicativas também deve-se usar práticas que garanta a qualidade da informação e, assim, com os dados fenotípicos e genotípicos de boa qualidade torna-se essencial a escolha de um bom modelo biométrico capaz de captar a estrutura genética envolvida no fenômeno estudado contemplando apectos relativos aos componentes da variação genética (aditividade, dominância e epistasia) e da interação entre genótipos e ambientes, dentre outros (PÉREZ; DE LOS CAMPOS, 2014).

Tendo em vista o envolvimento de matrizes de alta dimenssão, torna-se ainda mais evidente a necessidade de uma escolha bem fundamentada do modelo. Tais matrizes certamente, representam grande desafio, exigindo métodos de ajuste de modelo não tradicionais além de grande poder de processamento. Ao trabalhar com matrizes de alta dimensão, em seleção genômica, comumente são utilizados métodos bayesianos (MEUWISSEN; HAYES; GODDARD, 2001).

Ajustar modelos utilizando tais métodos, muitas vezs envolve um grande número de iterações para que a convergência do modelo seja atingida, assim, pequenas diferenças em termo de tempo computacional e métodos de aproximação adotados por diferentes ambientes de programação podem acarretar no fim do processo em diferenças na eficácia em termos de acurácia e eficiência com relação a poder de processamento. Assim utilizando as medidas, *acurácia seletiva e acurácia preditiva* para avaliar a acurácia e *velocidade de processamento* para avaliar o poder de processamento, foi comparado o desempenho dos pacotes BGLR-R e BGLR-Julia.

Nas Tabelas 1 e 2 verifica-se diferença entre os valores de acurácia seletiva encontrados utilizando o pacote BLGR, processado em diferentes ambientes de programação. De maneira geral, quando se processa dados em modelos menos

parametrizados e sem problemas de dimensionalidade e, ou, multicolinearidade, os parâmetros podem ser estimados por do método de quadrados mínimos, fornecendo assim uma única solução (RESENDE; SILVA; AZEVEDO, 2014). Entretanto, ao trabalhar com alta dimensionalidade e optarmos em ajustar modelos paramétricos tais como o Bayes A e Bayes B passamos a ter soluções aproximadas para os parâmetros, conduzindo a um infinito número de soluções (PÉREZ; DE LOS CAMPOS, 2014). Desta forma pode-se obter estimativas dos parâmetros distintas utilizando o mesmo modelo e consequentemente valores de acurácia distintos.

A acurácia seletiva apresentada na Tabela 1 e 2 representa a correlação entre o valor observado e o valor predito para tempo de floração nas plantas de arroz. Em um programa de melhoramento muitas vezes o principal interesse consiste em selecionar um grupo de indivíduos que possua maior potencial genético para a característica de interesse. Neste cenário, mesmo com um grande desvio do valor observado para o predito, se o modelo ajustado tiver um alto valor de acurácia seletiva o melhorista poderia utilizar os valores preditos como critério de seleção e obter sucesso na tomada de decisão. Desta forma dizer que um modelo possui alta acurácia seletiva equivale a dizer que os valores estimados e observados são altamente correlacionados.

Das Tabelas 1 e 2 é possível verificar que o BGLR-R apresentou melhores resultados que o BGLR-Julia em termos de acurácia seletiva quando se considerou todo o conjunto de informações do banco de dados molecular disponível. Esta superiordade persistiu quando foram considerado menores conjuntos de marcadores. Considerando apenas a validação, a diferença entre as acurácias seletivas nos dois ambientes é expressiva variando entre 12% (1205 marcadores) e 20% (16205 marcadores).

b. Acurácia preditiva

Em algumas situações, geralmente envolvendo melhoramento genético de animais de grande porte ou melhoramento vegetal de culuras perenes, o valor genético predito de cada indivíduo é informação importante, tendo em vista a potencialidade de uso comercial de cada indivíduo. Métodos que possibilitam a obtenção de estimativas de valores preditos próximos dos valores reais tem sido preferidos nesta situação e a informação sobre a eficácia deste método tem sido realizada por meio de sua acurácia preditiva.

Acurácia preditiva é dada pelo erro quadrático médio (EQM) do valor

observado e do valor predito para tempo de floração e suas estimativas estão apresentadas nas Tabelas 3 e 4 ao se fazer estudos com base nos modelos Bayes A e Bayes B.

Tabela 3. Acurácia preditiva (erro quadrático médio) obtida em análise de predição de valores genéticos de plantas de arroz, utilizando o modelo bayesiano *bayesA*, considerando diferentes conjuntos de informações moleculares

Número de Marcadores	R		Julia	
	Treinamento	Validação	Treinamento	Validação
26205	22.04	99.29	50.05	144.63
21205	24.37	99.60	45.95	157.57
16205	25.60	104.84	55.93	166.87
11205	42.07	113.57	67.04	146.04
6205	42.45	115.05	65.42	169.96
1205	63.58	122.58	82.67	155.55

A alta acurácia preditiva está relacionada a baixos valores do EQM e valores acurados nos garante sucesso não apenas no processo compartivo de seleção de indivíduos, como a acurácia seletiva, mas também garante bons resultados na predição de valores genotípicos para cada particular indivíduo. Assim, em um programas cujo exito em predizer o valor fenotípico do indivíduo tem principal foco, alta acurácia preditiva torna-se fundamental.

Novamente, verifica-se que o BGLR-R apresentou melhor desempenho. Nas Tabelas 3 e 4 é observado que ele apresentou melhores resultados que o BGLR-Julia em termos de acurácia preditiva em todos os cenários avaliados.

Tabela 4. Acurácia preditiva (erro quadrático médio) obtida em análise de predição de valores genéticos de plantas de arroz, utilizando o modelo *bayesB*, considerando diferentes conjuntos de informações moleculares

Número de Marcadores	R		Julia	
	Treinamento	Validação	Treinamento	Validação
26205	22,11	99,36	48,14	156,63
21205	23,65	99,92	35,75	161,4
16205	26,25	103,53	47,01	166,15
11205	38,59	111,52	60,93	180,67
6205	39,32	115,72	70,3	172,54
1205	56,73	122,42	82,2	158,64

c. Velocidade de processamento

O crescente volume de informações disponível aumenta a complexidade dos dados gerando desafios. Técnicas tradicionais de análise de dados apresentam limitações para big data, em relação a dados de alta dimensões (CHIAVEGATTO FILHO, 2015). Neste cenário escolher um ambiente de programação que apresente poder de processamento pode agregar para o pesquisador grande ganho em termos de tempo computacional gasto na análise (BEZANSON et al., 2012)

As estimativas apresentadas nas Tabelas 5 e 6, permite observar o desempenho, em termos de tempo de processamento, dos dois ambientes em estudo, aplicativo R e Julia, utilizando o pacote BGLR para dois modelos distintos (Bayes A e Bayes B).

Verifica-se que o BGLR-R, altamente otimizado em C, demonstrou maior poder de processamento, se mostrando mais eficiente em termos de tempo computacional do que o BGLR- Julia.

Tabela 5. Tempo de processamento (em minutos) em análise de predição de valores genéticos de plantas de arroz, utilizando o modelo bayesA, considerando diferentes conjuntos de informações moleculares

Número de Marcadores	Linguagem R	Linguagem Julia
26205	40,42	65,22
21205	28,75	52,25
16205	22,04	39,75
11205	15,3	31,6
6205	8,29	18,15
1205	1,43	2,77

Tabela 6. Tempo de processamento (em minutos) em análise de predição de valores genéticos de plantas de arroz, utilizando os modelos bayesB, considerando diferentes conjuntos de informações moleculares.

Número de Marcadores	Linguagem R	Linguagem Julia
26205	46,79	81,69
21205	37,89	58,82
16205	27,55	44,76
11205	18,46	33,23
6205	10,23	18,89
1205	2,08	3,63

Matosak, Medeiros e Santos (2017) compararam a performance dos ambientes de programação R e Julia quanto ao emprego de suas funções e em relação ao tempo necessário à execução de algoritmos usados no processamento digital de imagens orbitais. Os resultados obtidos mostram que com o aumento na dimensão dos dados de entrada a linguagem Julia apresenta menor tempo de execução. Porém, quando os dados de entrada não possuem grandes dimensões a linguagem R mostrou melhor desempenho.

Em contrapartida, este estudo mostou que a superioridade de desempenho do pacote BGLR no ambiente de programação R, tornou-se ainda mais evidente ao aumentar o número de marcadores. Tal fato é mostrado nas análises a seguir.

Considerando os resultados para o Bayes A, o modelo de regressão linear de

1° grau ajustado para o R foi: $T = -1{,}230 + 0{,}001503\,N$, com coeficiente de determinação $R^2 = 98{,}66\%$, tendo como variável resposta o tempo (T) gasto para o ajuste do modelo e variável preditora o número de marcadores (N) utilizado. Para a linguagem Julia, o modelo ajustado foi: $T = 1{,}853 + 0{,}002415\,N$ com $R^2 = 99{,}26\%$. Nas Tabelas 7 e 8 é apresentado o quadro das análises de variância.

Tabela 7. Análise de variância para a regressão tendo como variável resposta o tempo de processamento e como variável preditora o número de marcadores utilizados para o ajuste dos modelos no ambiente de programação R a partir do Bayes A.

Fonte	GL	SQ	QM	F	P
Regressão	1	988,655	988,655	368,35	0,000
Erro	4	10,736	2,684		
Total	5	999,391			

Tabela 8. Análise de variância para a regressão tendo como variável resposta o tempo de processamento e como variável preditora o número de marcadores utilizados para o ajuste dos modelos no ambiente de programação Julia a partir do Bayes A.

Fonte	GL	SQ	QM	F	P
Regressão	1	2552,50	2552,50	675,56	0,000
Erro	4	15,11	3,78		
Total	5	2567,62			

Os coeficientes de determinação mostram que os modelos foram bem ajustados (R^2 próximo de 100%) e pelos resultados obtidos nas análises de variância, constatatou-se a significância da regressão (P valor = 0). Além disto, levando em consideração os coeficientes angulares das equações, é observado que ao acrescentar um marcador, o tempo gasto para ajustar o modelo na linguagem Julia aumenta em 0,002415 minutos, enquanto no R o aumento é de 0,001503 minutos.

Para o Bayes B, o modelo ajustado no R foi: $T = -0{,}8842 + 0{,}001804\,N$ com $R^2 = 99{,}78\%$. Para a linguagem Julia, o modelo ajustado foi: $T = -0{,}679 + 0{,}002981\,N$ com $R^2 = 98{,}66\%$. Foi obtido um o bom ajustamento

dos modelos e as regressões foram significativas conforme observado nas análises de variância (Tabelas 9 e 10). Além disso, o acréscimo no tempo gasto para realizar a modelagem na linguagem Julia é maior que no R, ao aumentar o número de marcadores.

Tabela 9. Análise de variância para a regressão tendo como variável resposta o tempo de processamento e como variável preditora o número de marcadores utilizados para o ajuste dos modelos no ambiente de programação R a partir do Bayes B.

Fonte	GL	SQ	QM	F	P
Regressão	1	1423.09	1423.09	2316.37	0.000
Erro	4	2.46	0.61		
Total	5	1425.54			

Tabela 10. Análise de variância para a regressão tendo como variável resposta o tempo de processamento e como variável preditora o número de marcadores utilizados para o ajuste dos modelos no ambiente de programação Julia a partir do Bayes B.

Fonte	GL	SQ	QM	F	P
Regressão	1	3886.96	3886.96	370.05	0.000
Erro	4	42.02	10.50		
Total	5	3928.97			

4 Conclusão

Para o controle de qualidade dos marcadores as linguagens demostraram desempenho similar em termos de tempo de processamento. Ambas foram capazes de realizar o processo de filtragem dos dados de forma rápida e eficiente, o que ressalta o poder de processamento de ambas.

Para predições de valores genéticos utilizando os modelos de seleção disponíveis no pacote BGLR, sua versão para a linguagem Julia, se mostrou inferior em termos de tempo computacional, e também para acurácia seletiva e preditiva, ao ser comparado a sua versão desenvolvido para o R.

5 Referências Bibliográficas

ANDERSON, C. A. et al. Data quality control in genetic case-control association studies. **Nature Protocols**, v. 5, n. 9, p. 1564–1573, 2010.

BEZANSON, J. et al. Julia: A Fast Dynamic Language for Technical Computing. p. 1–27, 2012.

CAIXETA, E. T.; BORÉM, A. **Marcadores Moleculares**. Viçosa: UFV, 2006.

CAVALCANTI, J. J. V. et al. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, v. 34, n. 3, p. 840–846, 2012.

CHIAVEGATTO FILHO, A. D. P. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, v. 24, n. 2, p. 325–332, 2015.

FRANKE, B. et al. Statistical Inference, Learning and Models in Big Data. **International Statistical Review**, v. 84, n. 3, p. 371–389, 2016.

HEIDELBERGER, P.; WELCH, P. D. Simulation Run Length Control in the Presence of an Initial Transient. **Operations Research**, v. 31, n. 6, p. 1109–1144, 1983.

HUANG, X. et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. **Nature Genetics**, v. 44, n. 1, p. 32–39, 2012.

ILLUMINA INC. BovineHD Genotyping BeadChip. **Pub. No. 370-2010-018**, v. WG-450-100, p. 1–4, 2012.

IRRI - Research. Disponível em: <http://irri.org/our-work/research>. Acesso em: 30 jan. 2018.

LAURIE, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, v. 34, n. 6, p. 591–602, 2010.

MARTYN, A. et al. **Package ‘ coda ’R news**, 2016. Disponível em: <https://cran.r-project.org/web/packages/coda/coda.pdf>

MATOSAK, B. M.; MEDEIROS, N. DAS G.; SANTOS, G. R. DOS. Avaliação do desempenho dos ambientes r e julia com uso de operadores de realce e análise de imagens digitais. n. 2012, p. 24–26, 2017.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

MORAIS JÚNIOR, O. P. et al. **SELEÇÃO GENÔMICA AMPLA (GWS) APLICADA AO MELHORAMENTO POPULACIONAL**. [s.l.] UNIVERSIDADE FEDERAL DE GOIÁS, 2013.

PEIXOTO, L. DE A.; BHERING, L. L. **ABORDAGEM SOBRE MODELOS, COVARIÁVEIS E ACURÁCIA NA SELEÇÃO GENÔMICA**. [s.l.] Universidade Federal de Viçosa, 2016.

PEREZ-RODRIGUEZ, P.; DE LOS CAMPOS, G. **No Title**. Disponível em: <https://github.com/gdlc/BGLR.jl>.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

PINTOS, E. et al. **Modeling of RAFT Polymerization Processes Using an Efficient Monte Carlo Algorithm in Julia**. [s.l: s.n.]. v. 55

PLANT, S. et al. Genetic Control of Flowering Time in Rice , a NATURAL VARIATIONS : A NEW RESOURCE FOR. **Society**, v. 127, n. December, p. 1425–1429, 2001.

RESENDE, M. D. V. DE; SILVA, F. F. E; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa: Suprema, 2014.

TARANTO, F. et al. Genome-wide SNP discovery and population structure analysis in pepper (Capsicum annuum) using genotyping by sequencing. **BMC Genomics**, v. 17, n. 1, p. 943, 2016.

TEAM, R. C. The R Project for Statistical Computing. **Http://Www.R-Project.Org/**, 2013.

ZHAO, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature Communications**, v. 2, p. 467, 2011.