DOCUMENTOS

164

ISSN 1677-9274
Dezembro/2019

Genotipagem por meio de marcadores SNP em cevada (*Hordeum vulgare*): um estudo de caso em cultivares e populações resultantes de retrocruzamentos

Embrapa
Informática Agropecuária

ISSN 1677-9274
Dezembro/2019

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 164

Genotipagem por meio de marcadores SNP em cevada (*Hordeum vulgare*): um estudo de caso em cultivares e populações resultantes de retrocruzamentos

Maurício de Alvarenga Mudadu
Elene Yamazaki Lau
Jorge Fernando Pereira
Euclydes Minella

Autores

Embrapa Informática Agropecuária
Campinas, SP
2019

Exemplares desta publicação podem ser adquiridos na:

Embrapa Informática Agropecuária

Av. Dr. André Tosello, 209 - Cidade Universitária
Campinas, SP, Brasil
CEP. 13083-886
Fone: (19) 3211-5700
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Supervisão editorial
Kleber Xavier Sampaio de Souza

Revisão de texto
Edsel Rodrigues Teles

Normalização bibliográfica
Victor Paulo Marques Simão

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Felipe Prado Jaconi sob supervisão de Magda Cruciol

Foto da capa
Pixabay

1ª edição
Versão digital (2019)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Genotipagem por meio de marcadores SNP em cevada (Hordeum vulgare): um estudo de caso em cultivares e populações resultantes de retrocruzamentos / Maurício de Alvarenga Mudadu… [et al.]. – Campinas: Embrapa Informática Agropecuária, 2019.

PDF (21 p.) : il. – (Documentos / Embrapa Informática Agropecuária , ISSN 1677–9274 ; 164).

1. Marcador molecular. 2. Chip SNP. 3. Genotipagem. 4. Melhoramento genético. 5. Cevada. I. Mudadu, Maurício de Alvarenga. II. Título. III. Embrapa Informática Agropecuária. IV. Série.

CDD (21. ed.) 572.86

Victor Paulo Marques Simão (CRB-8/5139)

Autores

Maurício de Alvarenga Mudadu
Biólogo, doutor em Bioinformática
Pesquisador da Embrapa Informática Agropecuária, Campinas, SP

Elene Yamazaki Lau
Engenheira florestal, doutora em Genética e Melhoramento
Pesquisadora da Embrapa Trigo, Passo Fundo, RS

Jorge Fernando Pereira
Biólogo, doutor em Microbiologia Agrícola
Pesquisador da Embrapa Gado de Leite, Juiz de Fora, MG

Euclydes Minella
Agrônomo, PhD em Plant Breeding
Pesquisador da Embrapa Trigo, Passo Fundo, RS

Apresentação

A cevada, uma das primeiras espécies domesticadas pelo homem e um dos principais cereais produzidos no mundo, apresenta ampla gama de uso. Entretanto, o cultivo comercial de cevada no Brasil é exclusivamente destinado à produção de malte, para a fabricação de cerveja. Esse cultivo, primariamente realizado na região Sul, está associado à geração de milhões de empregos na cadeia produtiva e a um faturamento anual de aproximadamente 75 bilhões de reais do setor cervejeiro. Cerca de dois terços dos brasileiros preferem cerveja como bebida para comemorar os bons momentos, sendo o Brasil, em volume de consumo total, o terceiro maior mercado para cerveja no mundo.

A Embrapa possui papel significativo na geração de cultivares de cevada adaptadas às condições brasileiras. Em seus mais de 40 anos de atuação, o melhoramento realizado pela Embrapa, que trabalha em parceria com o setor privado, resultou em importantes cultivares que, atualmente, ocupam cerca de 90% da área plantada com cevada no Brasil. O programa de melhoramento de cevada da Embrapa e de seus parceiros ainda tem desafios importantes como, por exemplo, aumentar a produtividade e atender aos padrões de qualidade demandados pela indústria cervejeira. O momento é oportuno para que o melhoramento de cevada no Brasil utilize todas as ferramentas possíveis para superar seus desafios.

Dentre as ferramentas disponíveis, encontra-se o chip de marcadores moleculares do tipo *single nucleotide polymorphism* (SNP), que permite a análise de milhares de marcadores SNP espalhados no genoma da cevada. A utilização desses marcadores SNP pode auxiliar imensamente os melhoristas de cevada, bem como todos os interessados em estudos com essa espécie. Esses marcadores abrem possibilidades de estudos mais assertivos em mapeamento de QTLs, associação genômica, diversidade genética, incorporação de genes via cruzamento e retrocruzamentos e muitos outros. Dessa forma, é importante divulgar o conhecimento para obter e analisar dados de marcadores SNPs em cevada.

Este documento representa o primeiro passo no sentido de aumentar a utilização de marcadores SNP em cevada por grupos de pesquisa no Brasil. As formas de obtenção e análise dos dados são detalhadas utilizando-se estudo de caso sobre diversidade genética de cultivares e incorporação de um importante gene por meio de cruzamento e retrocruzamentos. Este relato aproveita a experiência de pesquisadores da Embrapa Trigo e da Embrapa Informática Agropecuária e abre a possibilidade para que milhares de marcadores SNP possam ser utilizados para diferentes finalidades, potencializando um melhoramento de cevada mais preciso e com maiores ganhos genéticos.

Silvia Maria Fonseca Silveira Massruhá

Chefe-geral

Embrapa Informática Agropecuária

Sumário

Introdução

A cevada é o quarto cereal mais importante do mundo. Essa espécie pode ser destinada para alimentação animal, para consumo humano na forma de bebidas e alimentos, e para a produção de medicamentos (De Mori; Minella, 2012; Estados Unidos, 2019). Em valores aproximados, 65% da cevada produzida no mundo são destinadas à alimentação animal, 33% à produção de bebidas e 2% à produção de alimentos a serem consumidos pelo homem (Baik; Ullrich, 2008). Apesar de sua importância mundial, a produção brasileira de cevada não atende à demanda de consumo interno. Nas últimas cinco safras (2014 a 2018), a área média destinada ao plantio de cevada no Brasil foi de cerca de 110 mil hectares, com produção média de aproximadamente 315 mil toneladas (CONAB, 2019). Adicionalmente, 75% da cevada produzida têm como destino a fabricação de malte e 95% do malte são destinados para fins cervejeiros (De Mori; Minella, 2012). Para atender ao consumo brasileiro de cerveja (57 litros per capita), cerca de 300 mil toneladas de grãos de cevada têm de ser importados juntamente com um expressivo volume de malte, extrato de malte e cerveja (De Mori; Minella, 2012). Futuros cenários de eventos extremos de seca e calor provavelmente diminuirão o rendimento de cevada em vários países, o que afetará a demanda global e o consumo de cerveja ao redor do mundo (Xie et al., 2018).

O programa de melhoramento genético de cevada da Embrapa, que funciona em parceria com o setor industrial, objetiva a obtenção de cultivares que atendam aos padrões de qualidade da indústria cervejeira (Minella, 2005). Além da qualidade, as características mais importantes para o melhoramento são alto rendimento e estabilidade da performance. Para tanto, a criação de cultivares para fins cervejeiros é baseada no desenvolvimento de germoplasma básico oriundo de diversos programas e bancos de germoplasma do mundo e na seleção em populações híbridas (melhoramento varietal) (Minella, 2005). Recentemente o uso de informações moleculares no programa de melhoramento genético de cevada da Embrapa tem aumentado principalmente com a análise da variabilidade genética de cultivares (Ferreira et al., 2016) e com a incorporação de marcadores moleculares para a tolerância ao alumínio (Ferreira et al., 2018).

Não apenas em cevada, como também no melhoramento de plantas de forma geral, o uso crescente de informações genômicas é uma das prerrogativas para aumentar a precisão e os ganhos genéticos do melhoramento em um contexto conhecido como melhoramento genético de nova geração, do inglês *next generation breeding* (Barabaschi et al., 2016). Dentre as ferramentas moleculares que podem ser utilizadas no melhoramento de nova geração em plantas com valor econômico, encontram-se os chips de marcadores genéticos ou *single nucleotide polymorphism* (SNP). Várias espécies de interesse econômico já apresentam chips comerciais com dezenas de milhares de marcadores disponíveis, tais como milho (Ganal et al., 2011), soja (Song et al., 2013), trigo (Wang et al., 2014) e café (Merot-L'Anthoene et al., 2019). A cevada também é uma das espécies para as quais chips de marcadores genéticos do tipo SNP estão disponíveis, como o *Illumina 50k iSelect,* que apresenta o diferencial de ter marcadores com posição genômica física acurada e anotação detalhada de genes (Bayer et al., 2017). Entretanto, o uso de informações originadas de chips de marcadores SNPs no programa de melhoramento de cevada da Embrapa ainda é incipiente.

Um chip de marcadores genéticos do tipo SNP possibilita genotipar amostras de forma menos custosa que o sequenciamento de genoma completo. A genotipagem por meio de SNP pode ser usada para verificar a similaridade genética entre indivíduos, auxiliar em programas de melhoramento genético na seleção assistida por marcadores ou mesmo em estudos de associação genômica ampla, ou GWAS (revisto por Mammadov et al., 2012). Em razão da importância dos marcadores SNP, o detalhamento de procedimentos para genotipagem e análise dos dados pode ser de grande

valia para diversos grupos de pesquisa. Assim, este documento tem como objetivo descrever com detalhes o processo para realizar a genotipagem de amostras de cevada usando o chip *Illumina 50k iSelect* e detalhar a metodologia de análise dos dados usada para obter a similaridade genética entre as amostras, de forma a verificar se os cruzamentos ocorreram a contento. Como estudo de caso, sete genótipos de cevada e plantas em quarta retrocruza, obtidas a partir do cruzamento de quatro cultivares e uma linhagem, foram genotipadas com o chip *Illumina 50k iSelect* e os dados analisados. Espera-se que grupos de pesquisa interessados em usar essa metodologia possam utilizar este documento norteador para seus experimentos e análises.

1. Seleção de amostras e cruzamentos

Quatro cultivares de cevada (Antarctica 01, BRS Cauê, BRS Itanema e MN 6021) e uma linhagem (Golden Promise-L5) foram utilizados para realizar cruzamentos. Antarctica 01, BRS Cauê, BRS Itanema, e MN 6021 (parentais recorrentes) foram utilizados como doadores de pólen em cruzamentos com Golden Promise-L5 (receptor de pólen e doador do gene *TaALMT1)*. As plantas originadas de cada cruzamento foram retrocruzadas com os parentais recorrentes quatro vezes (RC4). Nessa geração, as plantas foram autofecundadas e as gerações foram avançadas até F2RC4 ou F3RC4, sendo então selecionadas oito plantas de cada retrocruza. Assim, um total de 40 amostras foram utilizadas, sendo duas plantas da cultivar Golden Promise (Golden Promise-A e Golden Promise-B; sem o gene *TaALMT1*); uma de Antarctica 01, BRS Cauê, BRS Itanema, Dayton (controle com genoma menos relacionado) e MN 6021; uma de Golden Promise-L5; e as 32 amostras geradas por retrocruzamentos e autofecundações da linhagem Golden Promise-L5 com quatro parentais recorrentes, com a formação das seguintes “famílias”: Antarctica 01 x Golden Promise-L5 (família “AxL”), BRS Cauê x Golden Promise-L5 (família “CxL”); BRS Itanema x Golden Promise-L5 (família “IxL”) e MN 6021 x Golden Promise-L5 (família “MxL”) (Tabela 1).

Tabela 1. Identificação de amostras e famílias.

Identificação da família	Identificação da amostra
AxL	Antarctica_01
AxL	AxL_F2RC4-3-74-15
AxL	AxL_F3RC4-6-3-86-30
AxL	AxL_F2RC4-3-71-2
AxL	AxL_F3RC4-6-5-82-2
AxL	AxL_F3RC4-6-3-85-3
AxL	AxL_F3RC4-6-3-85-4
AxL	AxL_F3RC4-6-5-83-1
AxL	AxL_F3RC4-6-5-83-2
CxL	BRS_Cauê
CxL	CxL_F2RC4-3-75-3
CxL	CxL_F2RC4-3-76-45(2)
CxL	CxL_F2RC4-3-76-47
CxL	CxL_F2RC4-3-76-49
CxL	CxL_F3RC4-6-14-89-1
CxL	CxL_F3RC4-6-14-90-15
CxL	CxL_F3RC4-6-14-89-55
CxL	CxL_F3RC4-6-14-90-69
Dayton	Dayton
Golden_Promise_A	Golden_Promise_A
Golden_Promise_B	Golden_Promise_B
IxL	BRS_Itanema
IxL	IxL_F3RC4-1,3-6-108-257
IxL	IxL_F3RC4-1,3-6-108-259
IxL	IxL_F3RC4-1,3-6-108-260(2)
IxL	IxL_F3RC4-1,3-6-110-264
IxL	IxL_F3RC4-1,3-6-110-266
IxL	IxL_F3RC4-1,3-6-110-267
IxL	IxL_F3RC4-1,3-1-103-81
IxL	IxL_F3RC4-1,3-1-103-84
Golden_Promise_L5	Golden_Promise_L5
MxL	MN_6021
MxL	MxL_F3RC4-1,2,3-7-181-239
MxL	MxL_F3RC4-1,2,3-7-181-241
MxL	MxL_F3RC4-1,2,3-7-181-243
MxL	MxL_F3RC4-1,2,3-7-181-246
MxL	MxL_F2RC4-3-112-107
MxL	MxL_F2RC4-3-112-115
MxL	MxL_F3RC4-1,2,3-2-88-117
MxL	MxL_F3RC4-1,2,3-3-169-232

2. Extração de DNA

Para a extração de DNA, cerca de 100 mg de tecidos foliares jovens de cada um dos 40 materiais de cevada foram coletados, colocados em microtubo de 2 mL contendo três esferas de aço inox de 2,3 mm de diâmetro e congelados em nitrogênio líquido. Os microtubos foram agitados em macerador/homogeneizador de amostras “Mini-BeadBeater-96” (*BioSpec Products*) por dois minutos. Em seguida, o DNA foi extraído conforme as instruções do kit “DNeasy Plant Mini” (*Qiagen*), sendo eluídas em dois volumes de 100 µL de água ultrapura. Após a extração, as amostras foram quantificadas em gel de agarose 1% pela comparação com padrões de concentração de 50 e 100 ng de DNA

lâmbda, assim como utilizando o espectrofotômetro “Nanodrop” (*Eppendorf*). Duas réplicas contendo cerca de 300 ng de DNA de cada amostra foram depositadas em placa de 96 poços de 200 µL, secas no equipamento “Savant SPD121P SpeedVac Concentrator” (*Thermo Scientific*) e enviadas para serem analisadas quanto aos marcadores genéticos.

3. Genotipagem com chip de marcadores SNP

O chip *Illumina 50k iSelect* foi criado a partir do exoma de 170 acessos selecionados de cevada e possui por volta de 49 mil marcadores, dos quais 43.461 são funcionais (Bayer et al., 2017). As 40 amostras de cevada foram genotipadas no *James Hutton Institute*[1] e o resultado entregue em um arquivo texto em formato tabular. O arquivo continha informação de 40 amostras, com 42.799 marcadores por amostra, e estava formatado com amostras dispostas nas linhas e marcadores em colunas, sendo que a primeira linha continha os identificadores dos marcadores e a primeira coluna os identificadores das amostras. Os alelos estavam codificados por nucleotídeos (ACGT), sendo que, quando o genótipo era heterozigoto, uma barra (“/”) separava os alelos (ex.: “A/T”), do contrário, um único nucleotídeo significava genótipo homozigoto para o dado alelo. O formato do arquivo era compatível com o *software Flapjack*[2] (Milne et al., 2010).

Análises de dados e resultados

1. Softwares e sistemas utilizados

1.1 Sistema Operacional e Sistema Computacional

Foi utilizada uma máquina *desktop* comum, Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz, com 16 Gb de RAM e 1 Gb de armazenamento. O sistema operacional usado foi o Ubuntu 18.04.

1.2 Lista de softwares de terceiros

I. “Plink v1.9”[3] (manipulação e controle de qualidade de dados de genotipagem).

II. FlapJack v.1.19.03.18[4] (visualização e manipulação de dados de genotipagem)

III. Python 3.6.8 (default, Jan 14 2019, 11:02:34) (linguagem de programação)

IV. “R version 3.4.4[5] (2018-03-15)” (linguagem de programação)

- “Bioconductor”, pacote “SnpStats”.[6]

V. Rstudio Version 1.2.1335[7]

1 https://ics.hutton.ac.uk

2 Disponível em: <http://flapjack.hutton.ac.uk/en/latest/import_data.html#importing-data>.

3 Disponível em: <https://www.cog-genomics.org/plink2>.

4 Disponível em: <https://ics.hutton.ac.uk/flapjack/>.

5 Disponível em: <https://www.r-project.org/>.

6 Disponível em: <http://bioconductor.org/packages/release/bioc/html/snpStats.html>.

7 Disponível em: <https://www.rstudio.com/>.

1.3 *Scripts* desenvolvidos *"in house"*

Os *scripts* desenvolvidos para análise de dados estão listados a seguir:

I. *Scripts* na linguagem Python 3.6 com finalidades diversas foram versionados no sistema GitLab da Embrapa Informática Agropecuária e estão disponíveis a interessados:

i. "decode.py" - destinado a decodificar as amostras e convertê-los nos identificadores reais dos acessos de cevada utilizados.

ii. "barley_genotype_to_flapjack.py"[8]– destinado a filtrar os marcadores do arquivo original cujos identificadores não estavam presentes no arquivo com informações de marcadores.

iii. "barley_genotype_to_ped.py" - destinado a transformar os dados no formato original para o formato ".ped" (*pedigree*, ou *linkage format)*; para mais detalhes do formato ver informações na página do Broad Institute[9]

iv. "recode_01.py" - destinado a formatar os dados de genotipagem para gerar a matriz de similaridade no R. O script codifica os genótipos "1 1" em "0" (homozigoto para o alelo 1); "1 2" ou "2 1" em 1 (heterozigoto); "2 2" em 2 (homozigoto para o alelo2) e mantém o caractere "-" para os genótipos faltantes.

2. Formatação dos dados

As amostras foram enviadas para genotipagem com identificadores codificados e foi necessário decodificá-las (Figura 1.i). O *script* "decode.py" foi usado para esse intento, tendo como entradas o arquivo texto ("barley.ids"), que continha um dicionário dos nomes codificados e decodificados das amostras, e o arquivo de genotipagem "embrapa.txt". Como saída, foi gerado o arquivo "embrapa.txt.decoded.txt". Esse arquivo foi então filtrado (Figura 1.ii) com o script "barley_genotype_to_flapjack.py" para confirmar os identificadores utilizados no arquivo de genotipagem cruzando com o arquivo de informação dos SNP do chip (arquivo "snp_effects.txt") baixado do *James Hutton Institute*. Foram confirmados 41.306 marcadores, mantidos nos arquivos "genotypes.map.flapjack.txt", "genotypes.ped.flapjack.txt" para uso no *software FlapJack.*

```
i)python3 decode.py --dictionary barley.ids --genotypes embrapa.txt

ii)python3 barley_genotype_to_flapjack.py --map_file snp_effects.txt
 --genotype_file embrapa.txt.decoded.txt --verbose
```

Figura 1. Comandos em um shell do Linux para formatar os dados.

3. Controle de qualidade (QC)

Os dados do arquivo "embrapa.decoded.txt" foram filtrados usando o arquivo "snp_effects.txt" e transformados para o formato *pedigree* (Figura 2.i), gerando os arquivos "genotypes.ped" e "genotypes.map" com dados de 41.306 marcadores.

Para executar o controle de qualidade (QC) nos dados de genotipagem, o software "Plink v1.9" (Chang et al., 2015) foi utilizado (Figura 2.ii). Foram utilizados os seguintes filtros: *call rate* por

8 Disponível em: <https://ics.hutton.ac.uk/50k/resources/snp_effects.txt>.

9 Disponível em: <https://www.broadinstitute.org/science/programs/medical-and-population-genetics/haploview/input-file-formats-0>.

amostra (<90%), *call rate* por marcadores (<1e-12%) e MAF (*minor allele frequency*) (<1e-12%). Respectivamente, as *flags* de filtro de qualidade para o "Plink v1.9" foram: "--mind 0.1 --geno 0 --maf 1e-12". Os filtros de *call rate* por marcadores e MAF foram bastante estringentes, já que o número de marcadores era abundante e o objetivo seria estudar a similaridade entre as amostras. Dessa forma, marcadores que apresentaram qualquer falha de genotipagem ou marcadores não informativos (monomórficos) foram eliminados. Após o QC, todas as 40 amostras foram mantidas, evidenciando a boa qualidade da genotipagem. Foram eliminados 159 marcadores em razão do *call rate* e 12.676 removidos em razão do filtro de MAF. No total, os filtros de QC eliminaram em torno de 30% dos marcadores de cada cromossomo de forma aparentemente uniforme (variando entre 44% e 29%). A maioria dos marcadores eliminados foi não informativa (MAF igual a zero), o que demonstra que não houve viés nos filtros. Restaram então 28.471 marcadores e 40 amostras, que foram usados nas análises subsequentes. Foram gerados os arquivos transpostos "genotypes_QC.tped", "genotypes_QC.tfam" e os arquivos binários "genotypes_QC.bed", "genotypes_QC.bim" e "genotypes_QC.fam". Os genótipos do arquivo "genotypes_QC.tped" foram formatados para os códigos "0, 1 e 2" como descrito anteriormente (Figura 2.iii). Foram gerados os arquivos "genotypes_QC.tped.converted.tped" e "genotypes_QC.converted.tmap".

As informações de famílias e identificadores de amostras contidas na Tabela 1 foram passadas para um arquivo tabular "genotypes_QC.tped.converted.tfam" que será usado nas análises a seguir.

```
i) python3 barley_genotype_to_ped.py --map_file snp_effects.txt  --genotype_file embrapa.txt.decoded.txt --verbose

ii) plink --file genotypes --make-bed --chr-set 7 --allow-extra-chr --recode 12 transpose --no-fid --no-parents --no-sex
--no-pheno --out genotypes_QC --geno 1e-12 --mind 0.1 --maf 1e-12 --missing-genotype "-"

iii)python3 recode_01.py --ped genotypes_QC.tped
```

Figura 2. Comandos no shell do Linux para formatar dados para o formato .ped e filtrar dados usando um QC.

4. Similaridade e cluster hierárquico

Os arquivos binários gerados pelo Plink foram usados como entrada para o R, pacote "snpStats" (Clayton, 2019). Foi calculada a matriz de distância e similaridade usando o método de identidade por estado IBS (Figura 3). A matriz de distância foi usada para cálculo de um *cluster* hierárquico pelo método *Ward.D2* (Figura 4).

```
library(snpStats)
sample <- read.plink("genotypes_QC.bed", "genotypes_QC.bim", "genotypes_QC.fam")
ibs <- ibsCount(sample$genotypes)
ibsD <- ibsDist(ibs) # matriz de distância
hc2= hclust(ibsD, method = "ward.D2") # cluster hierárquico
tiff(file="hclust.tiff",width=1920, height=1080, units = "px", pointsize=20,
compression=c("none"), bg="white")
plot(hc2)
dev.off()
ibsS <- as.matrix(1-ibsD)# similarity matrix
# similarity between the same individuals is 0, but should be 1
ibsS <- ifelse(ibsS==0.0000000,1.0000000,ibsS)
# Similarity table
write.table(ibsS[,c("Golden_Promise_L5","Antarctica_01","BRS_Cauê","BRS_Itanema","MN_6021")],sep="\t")'
```

Figura 3. Comandos no R para cálculo das matrizes de distância, similaridade e cluster hierárquico.

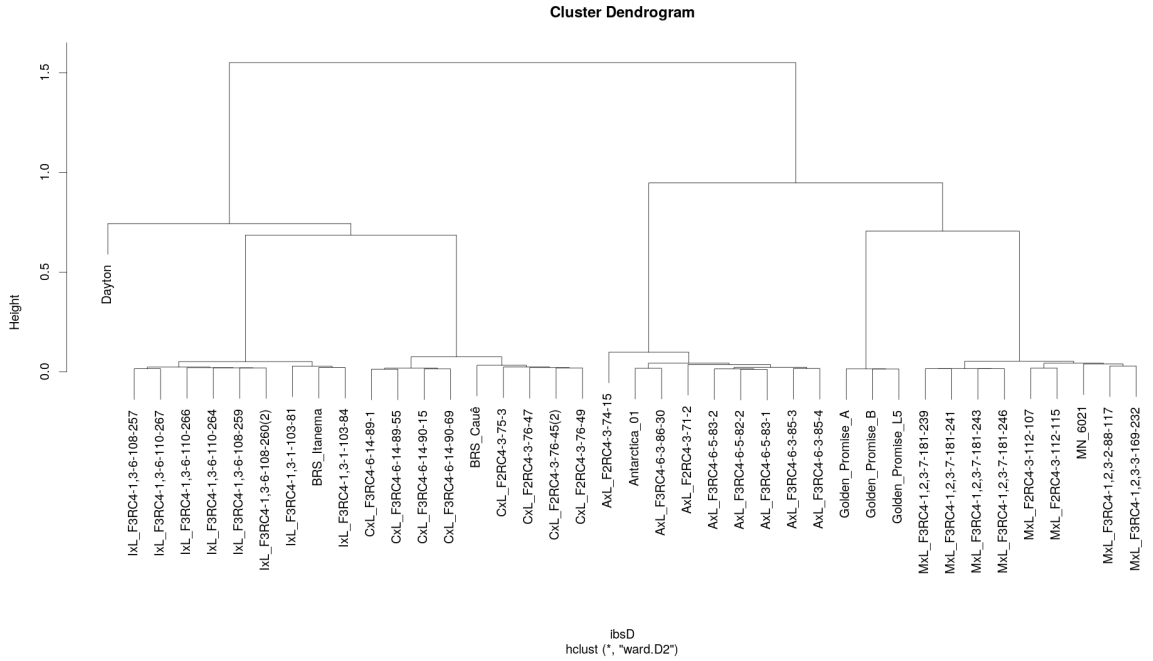

Figura 4. Cluster hierárquico.

5. *Heatmap* e PCA

Os arquivos "genotypes_QC.tped.converted.tped" e "genotypes_QC.converted.tmap" foram usados para construir uma matriz de relacionamento genômico (Tabela 2) de acordo com a fórmula de VanRaden (VanRaden, 2008; Gondro et al., 2013) usando a linguagem R (Figura 5). Com essa matriz é possível plotar a similaridade entre as amostras usando um gráfico do tipo *heatmap* (Figura 6), no qual cores mais claras evidenciam as amostras menos similares entre si e cores mais escuras o inverso. Também foi possível usar a GRM para calcular um PCA (Figura 7) e evidenciar o distanciamento das famílias.

Tabela 2. Matriz de similaridade entre amostras. Valores acima de 0,9 estão marcados.

	Golden_Promise_L5	*Antarctica_01*	*BRS_Cauê*	*BRS_Itanema*	*MN_6021*
Antarctica_01	0,6265	1,0000	0,5995	0,5855	0,6784
AxL_F2RC4-3-71-2	0,6484	0,9607	0,6041	0,5925	0,6819
AxL_F2RC4-3-74-15	0,6709	0,9335	0,5755	0,5696	0,6460
AxL_F3RC4-6-3-85-3	0,6484	0,9638	0,5943	0,5851	0,6728
AxL_F3RC4-6-3-85-4	0,6459	0,9671	0,5973	0,5856	0,6759
AxL_F3RC4-6-3-86-30	0,6311	0,9826	0,5995	0,5851	0,6780
AxL_F3RC4-6-5-82-2	0,6420	0,9718	0,6009	0,5862	0,6781
AxL_F3RC4-6-5-83-1	0,6426	0,9702	0,6005	0,5860	0,6786
AxL_F3RC4-6-5-83-2	0,6415	0,9710	0,6016	0,5867	0,6796
BRS_Cauê	0,6017	0,5995	1,0000	0,7814	0,5862
CxL_F2RC4-3-75-3	0,6247	0,5943	0,9638	0,7662	0,5881
CxL_F2RC4-3-76-45(2)	0,6118	0,5948	0,9763	0,7764	0,5813
CxL_F2RC4-3-76-47	0,6160	0,5941	0,9722	0,7727	0,5841
CxL_F2RC4-3-76-49	0,6153	0,5902	0,9726	0,7745	0,5808
CxL_F3RC4-6-14-89-1	0,6168	0,5977	0,9712	0,7666	0,5908
CxL_F3RC4-6-14-89-55	0,6165	0,5976	0,9712	0,7664	0,5908
CxL_F3RC4-6-14-90-15	0,6141	0,5969	0,9736	0,7686	0,5894
CxL_F3RC4-6-14-90-69	0,6147	0,5963	0,9732	0,7683	0,5908
Dayton	0,4048	0,4139	0,4436	0,4521	0,3922
Golden_Promise_A	0,9855	0,6268	0,6022	0,5857	0,6520
Golden_Promise_B	0,9860	0,6263	0,6018	0,5855	0,6520
BRS_Itanema	0,5854	0,5855	0,7814	1,0000	0,5651
IxL_F3RC4-1,3-1-103-81	0,5969	0,5840	0,7756	0,9724	0,5648
IxL_F3RC4-1,3-1-103-84	0,5931	0,5852	0,7789	0,9782	0,5661
IxL_F3RC4-1,3-6-108-257	0,6040	0,5871	0,7759	0,9670	0,5655
IxL_F3RC4-1,3-6-108-259	0,5990	0,5870	0,7779	0,9724	0,5666
IxL_F3RC4-1,3-6-108-260(2)	0,5982	0,5863	0,7774	0,9726	0,5662
IxL_F3RC4-1,3-6-110-264	0,5969	0,5879	0,7801	0,9738	0,5679
IxL_F3RC4-1,3-6-110-266	0,5980	0,5878	0,7783	0,9731	0,5657
IxL_F3RC4-1,3-6-110-267	0,6037	0,5884	0,7767	0,9684	0,5653
Golden_Promise_L5	1,0000	0,6265	0,6017	0,5854	0,6522
MN_6021	0,6522	0,6784	0,5862	0,5651	1,0000
MxL_F2RC4-3-112-107	0,6641	0,6791	0,5803	0,5584	0,9625
MxL_F2RC4-3-112-115	0,6685	0,6768	0,5792	0,5567	0,9579
MxL_F3RC4-1,2,3-2-88-117	0,6664	0,6794	0,5793	0,5577	0,9593
MxL_F3RC4-1,2,3-3-169-232	0,6603	0,6817	0,5847	0,5663	0,9669
MxL_F3RC4-1,2,3-7-181-239	0,6715	0,6731	0,5692	0,5489	0,9567
MxL_F3RC4-1,2,3-7-181-241	0,6710	0,6734	0,5699	0,5496	0,9557
MxL_F3RC4-1,2,3-7-181-243	0,6723	0,6720	0,5697	0,5492	0,9545
MxL_F3RC4-1,2,3-7-181-246	0,6718	0,6729	0,5691	0,5496	0,9552

```
library(dichromat)
library(gplots)
samples_snps=read.table("genotypes_QC.tped.converted.tped",header=F,sep=" ", na.strings="-")
samples_info=read.table("genotypes_QC.tped.converted.tfam",header=F,sep=" ")
map=read.table("genotypes_QC.tped.converted.tmap",header=F,sep="\t")
M=as.matrix(samples_snps, na.strings="-", stringsAsFactors=FALSE)
p=apply (M,1,function(x) sum(x)/(length(x)*2))
M=M-1
P=2*(p-0.5)
Z=M-P
ZtZ = t(Z) %*% Z
d=2*sum(p*(1-p))
G=ZtZ/d # GRM
SVD=svd(G)
colnames(SVD$v) <- samples_info$V2
rownames(SVD$v) <- samples_info$V2
colnames(G) <- samples_info$V2
rownames(G) <- samples_info$V2
samples_pc1 <- data.frame((SVD$v)[,1])
samples_pc2 <- data.frame((SVD$v)[,2])
colnames(samples_pc1) <- c("pc1")
colnames(samples_pc2) <- c("pc2")
samples_labels_pos <- cbind(samples_pc2, (-1*samples_pc1))
samples_labels_txt <- colnames(SVD$v)
samples_order = as.data.frame(colnames(SVD$v))
colnames(samples_order)=c("id")
### put family info
family_snps <- data.frame(colnames(SVD$v))
colnames(family_snps) <- c("id")
family_snps$famid <- NULL
family_snps_order <- merge( family_samples, family_snps, by=c("id"))
samples_order_family <- family_snps_order[samples_order$id,]
# Set colors by family
famcolschema <- rainbow(length(unique(samples_order_family$famid)))
Col_tag_famid <- famcolschema[samples_order_family$famid]
# Heatmap
tiff(file="barley.converted.heatmap.tiff",width=1920, height=1080, units = "px", pointsize = 20, compression=c("none"), bg="white")
heatmap(G, symm=T, col=gray.colors(16,start=0,end=1), margins=c(12,9))
dev.off()
#PCA
tiff(file="barley.pca.tiff",width=1920, height=1080, units = "px", pointsize = 20, compression=c("none"), bg="white")
plot(SVD$v[,1],-1*SVD$v[,2], cex.main=0.9, main="singular value decomposition", xlab="PC1", ylab="PC2", col=Col_tag_famid, pch=20, cex=1.5)
legend("topright", as.character(unique(samples_order_family$famid)), fil=c("#80FF00FF","#FF0000FF", "#FFBF00FF", "#0040FFFF", "#FF00BFFF",
"#00FF40FF", "#00FFFFFF", "#8000FFFF"), cex=1)
dev.off()
```

Figura 5. Comandos no R para cálculo e construção do PCA e *Heatmap*.

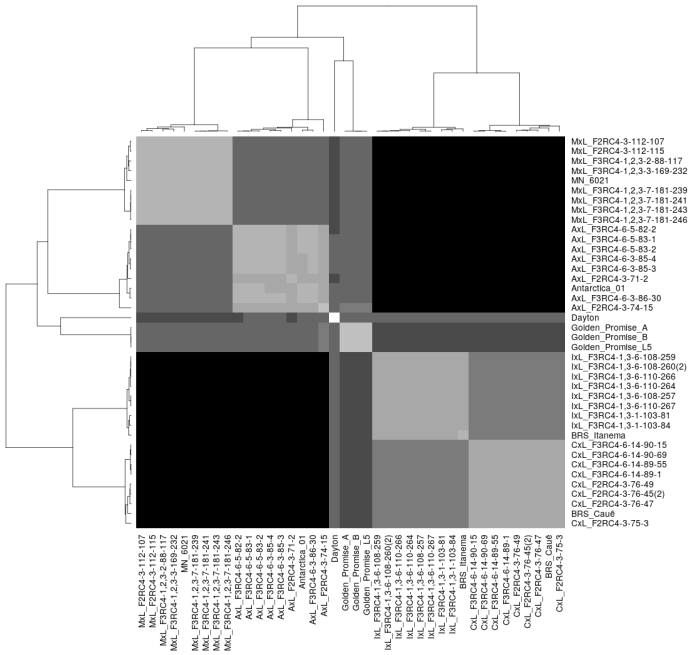

Figura 6. *Heatmap* e cluster hierárquico. Tons mais escuros representam menor similaridade.

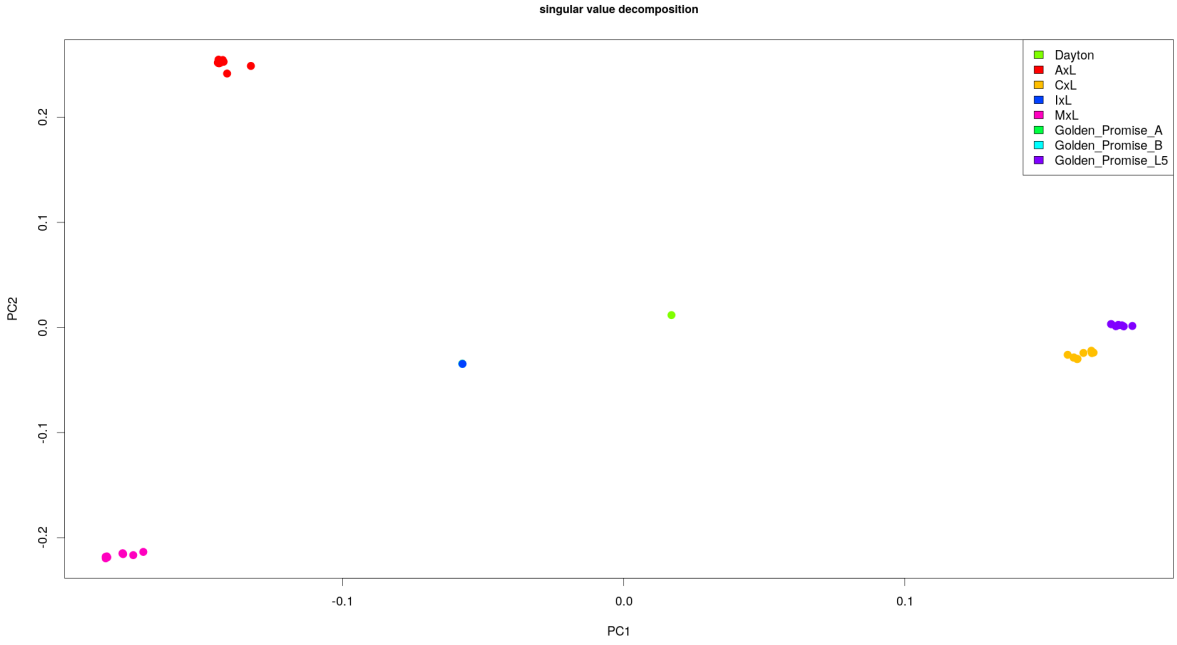

Figura 7. PCA com as cores evidenciando os diferentes indivíduos e famílias.

Discussão

Os resultados comprovaram que os cruzamentos ocorreram a contento, tendo em vista a similaridade entre indivíduos relacionados (Tabela 2), as matrizes de distância e a similaridade que podem ser visualizadas pelo *cluster* hierárquico, *heatmap* (Figuras 4 e 6) e também pelo PCA que separou as amostras como previsto (Figura 7).

Teoricamente, é esperada uma recuperação de 96,87% do genoma do parental recorrente em RC4 (Mesquita, 2002). Dentre as amostras analisadas, 78,13% apresentam mais de 96,00% do genoma

recorrente e 96,88% apresentam mais de 95,00%. Apenas uma amostra da família AxL apresentou percentual menor do genoma recorrente, sendo de 93,33%, e outras seis amostras da família MxL apresentaram entre 95,45% e 95,93% (Tabela 2, marcado em cores). Os dois acessos de Golden Promise e Golden Promise-L5 foram muito similares entre si e Dayton, com menor similaridade em relação a todas as amostras, variando de 39,22% a 45,21% (Tabela 2). Este resultado era esperado, visto que análises com marcadores microssatélites demonstraram que as cultivares brasileiras (Antarctica 01, BRS Cauê e BRS Itanema) e argentina (MN 6021) são mais similares entre si quando comparadas a Golden Promise e Dayton, que são cultivares obtidas na Inglaterra e nos Estados Unidos, respectivamente (Ferreira et al., 2016). De forma geral, genótipos brasileiros de cevada são mais similares entre si, o que indica que o uso de genótipos mais dissimilares pode ser interessante para programas de melhoramento de cevada no Brasil explorarem os benefícios de novos alelos (Ferreira et al., 2018). Os marcadores SNPs da plataforma *Illumina 50k iSelect* e o método de análise aqui descrito podem facilitar a avaliação da similaridade de genótipos utilizados por programas de melhoramento.

O *cluster* hierárquico e *heatmap* permitiram visualizar a maior proximidade entre as famílias CxL e IxL e entre as famílias AxL e MxL (Figuras 4 e 6). O PCA (Figura 7) confirmou os resultados das Figuras 4 e 6 por meio dos componentes principais (PC1 e PC2). As cultivares BRS Cauê e BRS Itanema apresentam em seu pedigree um parental comum (BRS 195), o que explica o padrão de similaridade encontrado.

Portanto, confirmou-se a efetividade do uso dos marcadores SNPs da plataforma *Illumina 50k iSelect* e do método de análise para caracterizar essas amostras resultantes de retrocruzamentos.

Os métodos de análise utilizados se mostraram efetivos, sendo que a maioria das figuras e análises foram feitas com métodos e bibliotecas de análise já bem estabelecidos, como pacotes do *Bioconductor/R, Plink e FlapJack,* sendo todos *softwares* livres. Houve pouca necessidade de intervenção bioinformática nos dados, sendo que apenas alguns *scripts* simples tiveram de ser construídos para adequar e converter os dados aos *softwares* de análise existentes.

Conclusão

O chip *Illumina iSelect 50k* para cevada é uma ferramenta genômica que traz versatilidade às análises genéticas e ao melhoramento de cevada e pode gerar resultados de forma bastante rápida, acurada e pouco custosa. Espera-se que o compartilhamento dessa metodologia experimental e de análise descrita neste documento possa servir para impulsionar o uso dessa e outras ferramentas em estudos genômicos e no melhoramento de cevada.

Agradecimentos

Os autores agradecem aos pesquisadores Peter Ryan e Manny Delhaize do *CSIRO Agriculture and Food*, Austrália, pelo envio de sementes da linhagem Golden Promise-L5; às técnicas da Embrapa Trigo, Andréa Morás e Lucimere de Fátima Morelo, pela condução das plantas e análises laboratoriais; ao assistente Ademir Vicari, pela realização dos cruzamentos; aos estudantes Gizele Carla Rogalsky, Natália Balbinott e Eduardo André Roesler; e a outros empregados da Embrapa Trigo pelos auxílios eventuais.

Referências

BAIK, B. K.; ULLRICH, S. E. Barley for food: characteristics, improvement, and renewed interest. **Journal of Cereal Science**, v. 48, n. 2, p. 233-242, 2008.

BARABASCHI, D.; TONDELLI, A.; DESIDERIO, F.; VOLANTE, A.; VACCINO, P.; VALÈ, G.; CATTIVELLI, L. Next generation breeding. **Plant Science**, v. 242, p. 3-13, 2016.

BAYER, M. M.; RAPAZOTE-FLORES, P.; GANAL, M.; HEDLEY, P. E.; MACAULAY, M.; PLIESKE, J.; RAMSAY, L.; RUSSELL, J.; SHAW, P. D.; THOMAS, W.; WAUGH, R. Development and evaluation of a barley 50k iSelect SNP array. **Frontiers in Plant Science**, v. 8, Oct. 2017. Article 1792. DOI: 10.3389/fpls.2017.01792.

CHANG, C. C.; CHOW, C. C.; TELLIER, L. C.; VATTIKUTI, S.; PURCELL, S. M.; LEE, J. J. Second-generation PLINK: rising to the challenge of larger and /richer datasets. **GigaScience**, v. 4, n. 1, 2015. Article 7. DOI :10.1186/s13742-015-0047-8.

CLAYTON, D. **SnpStats**: SnpMatrix and XSnpMatrix classes and methods. R package version 1.28.0. 2015. Disponível em: <https://rdrr.io/bioc/snpStats/>. Acesso em: 10 set. 2019.

CONAB. **Série histórica da área, produtividade e produção de cevada.** Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras/item/7681-cevada> Acesso em: 1 jun. 2019

DE MORI, C.; MINELLA, E. **Aspectos econômicos e conjunturais da cultura da cevada.** Passo Fundo: Embrapa Trigo, 2012. 28 p. (Embrapa Trigo. Documentos online, 139).

ESTADOS UNIDOS. Department of Agriculture. **World agricultural production**. Washington, DC, 2019. (Circular series. WAP, 10-19). Disponível em: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf>. Acesso em: 26 jun. 2019.

FERREIRA, J. R.; MINELLA, E.; DELATORRE, C. A.; DELHAIZE, E.; RYAN, P. R.; PEREIRA, J. F. Conventional and transgenic strategies to enhance the acid soil tolerance of barley. **Molecular Breeding**, v. 38, n. 1, p. 1-11,2018. Article 12. DOI: 10.1007/s11032-017-0769-7.

FERREIRA, J. R.; PEREIRA, J. F.; TURCHETTO, C.; MINELLA, E.; CONSOLI, L.; DELATORRE, C. A. Assessment of genetic diversity in Brazilian barley using SSR markers. **Genetics and Molecular Biology**, v. 39 n. 1, p. 86-96, 2016. DOI: 10.1590/1678-4685-GMB-2015-0148.

GANAL, M. W.; DURSTEWITZ, G.; POLLEY, A.; BÉRARD, A.; BUCKLER, E. S.; CHARCOSSET, A.; CLARKE, J. D.; GRANER, E.-V.; HANSEN, M.; JOETS,J.; LE PASLIER, M.-C.; MCMULLEN, M. D.; MONTALENT, P.; ROSE, M.; SCHÖN, C.-C.; SUN. Q.; WALTER, H.; MARTIN, O. C.; FALQUE, M. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. **PLoS ONE**, v. 6, n. 12, e28334, 2011. DOI: doi.org/10.1371/journal.pone.0028334

GONDRO, C.; WERF, J. van der; Hayes, B. (Ed.). **Genome-Wide Association Studies and Genomic Prediction.** Totowa: Humana, 2013. 566 p.

MAMMADOV, J.; AGGARWAL, R.; BUYYARAPU, R.; KUMPATLA, S. SNP markers and Their Impact on Plant Breeding. **International Journal of Plant Genomics**, v. 2012, p. 1–11, 2012. Article: 728398. DOI: 10.1155/2012/728398.

MEROT-L'ANTHOENE, V.; TOURNEBIZE, R.; DARRACQ, O.; RATTINA, V.; LEPELLEY, M.; BELLANGER, L.; TRANCHANT-DUBREUIL, C.; COULÉE, M.; PÉGARD, M.; METAIRON, S.; FOURNIER, C.; STOFFELEN, P.; JANSSENS, S. B.; KIWUKA, C.; MUSOLI, P.; SUMIRAT, U.; LEGNATÉ, H.; KAMBALE, J.; COSTA NETO, J. F. DA; REVEL, C.; KOCHKO, A. DE; DESCOMBES, P.; CROUZILLAT, D.; PONCET, V. Development and evaluation of a

genome-wide Coffee 8.5 K SNP array and its application for high-density genetic mapping and for investigating the origin of Coffea arabica L. **Plant Biotechnology Journal,** v. 17, n. 7, p. 1418-1430, 2019. DOI: 10.1111/pbi.13066

MESQUITA, A. G. G. **Retrocruzamento assistido por marcadores SSRs em milho.** 2002. 69 f. Tese (Doutorado) - Universidade Federal de Lavras, Lavras.

MILNE, I.; SHAW, P.; STEPHEN, G.; BAYER, M.; CARDLE, L.; THOMAS, W. T. B.; FLAVELL, A. J.; MARSHALL, D. Flapjack--graphical genotype visualization. **Bioinformatics**, v. 26, n. 24, p. 3133–3134, 2010. https://doi.org/10.1093/bioinformatics/btq580

MINELLA, E. Melhoramento de cevada. In: Borém, A. (Ed.). **Melhoramento de espécies cultivadas.** 2. ed. Viçosa, MG: Editora UFV, 2005. p. 275-299.

SONG, Q.; HYTEN, D. L.; JIA, G.; QUIGLEY, C. V.; FICKUS, E. W.; NELSON, R. L.; CREGAN, P. B. Development and evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE**, v. 8, n. 1, e54985, 2013. DOI :10.1371/journal.pone.0054985.

VanRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008. DOI: 10.3168/jds.2007-0980.

WANG, S.; WONG, D.; FORREST, K.; ALLEN, A.; CHAO, S.; HUANG, B. E.; MACCAFERRI, M.; SALVI, S.; MILNER, S. G.; CATTIVELLI, L.; MASTRANGELO, A. M.; WHAN, A.; STEPHEN, S.; BARKER, G.; WIESEKE, R.; PLIESKE, J.; LILLEMO, M.; MATHER, D.; APPELS, R.; DOLFERUS, R.; BROWN-GUEDIRA, G.; KOROL, A.; AKHUNOVA, A. R.; FEUILLET, C.; SALSE, J.; MORGANTE, M.; POZNIAK, C.; LUO, M.-C.; DVORAK, J.; MORELL, M.; DUBCOVSKY, J.; GANAL, M.; TUBEROSA, R.; LAWLEY, C.; MIKOULITCH, I.; CAVANAGH, C.; EDWARDS, K. J.; HAYDEN, M.; AKHUNOV, E. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. **Plant Biotechnology Journal,** v. 12, n. 6, p. 787-796, 2014.

XIE, W.; XIONG, W.; PAN, J.; ALI, T.; CUI, Q.; GUAN, D. ;MENG, J.; MUELLER, N. D., LIN, E.; DAVIS, S. J. Decreases in global beer supply due to extreme drought and heat. **Nature Plants,** v. 4, n. 11, p. 964-973, 2018.

Embrapa

Informática Agropecuária

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

CGPE 15550