

FERNANDA FÁTIMA CANIATO

**ESTRUTURA DE POPULAÇÃO E MAPEAMENTO ASSOCIATIVO PARA
TOLERÂNCIA AO ALUMÍNIO EM
REGIÕES CANDIDATAS DO LOCO Alt_{SB} DE SORGO**

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2009

**Ficha catalográfica preparada pela Seção de Catalogação e
Classificação da Biblioteca Central da UFV**

T

C223e
2009

Caniato, Fernanda Fátima, 1975-

Estrutura de população e mapeamento associativo para tolerância ao alumínio em regiões candidatas do loco *Alt_{SB}* de sorgo / Fernanda Fátima Caniato. – Viçosa, MG, 2009. vii, 116f.: il. (algumas col.) ; 29cm.

Orientador: Aluizio Borém.

Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Sorgo - Melhoramento genético. 2. Genética de populações. 3. Sorgo - Efeito do alumínio. I. Universidade Federal de Viçosa. II. Título.

CDD 22.ed. 633.1742

FERNANDA FÁTIMA CANIATO

**ESTRUTURA DE POPULAÇÃO E MAPEAMENTO ASSOCIATIVO PARA
TOLERÂNCIA AO ALUMÍNIO EM
REGIÕES CANDIDATAS DO LOCO Alt_{SB} DE SORGO**

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 27 de maio de 2009

Dra. Cláudia Teixeira Guimarães
(Co-orientadora)

Prof. Antônio Augusto Franco Garcia

Prof. José Marcelo Soriano Viana

Prof. Cosme Damião Cruz

Dr. Jurandir Vieira de Magalhães
Presidente da Banca

AGRADECIMENTOS

Agradeço à Universidade Federal de Viçosa pela oportunidade de realizar o curso de Doutorado em Genética e Melhoramento

À Fapemig, CNPq e CAPES pela concessão de bolsa de estudos.

À EMBRAPA Milho e Sorgo, pela oportunidade de conduzir os experimentos nesta instituição.

Ao Professor Aluizio Borém pela orientação acadêmica na Universidade Federal de Viçosa.

Ao Dr. Jurandir Vieira de Magalhães pela orientação na pesquisa conduzida na Embrapa Milho e Sorgo.

À Dra. Cláudia Teixeira Guimarães, pela co-orientação na pesquisa conduzida na Embrapa Milho e Sorgo.

Ao Dr. Robert Eugene Schafertt por toda a sua ajuda e suporte ao longo de todo o projeto.

A todos os funcionários e estagiários da Embrapa Milho e Sorgo, em especial aqueles do Núcleo de Biologia Aplicada que de alguma maneira tornaram possível a realização deste trabalho.

CONTEÚDO

RESUMO	iv
ABSTRACT	vi
Introdução Geral.....	1
1. Importância do sorgo.....	1
2. Efeito da Toxidez do Alumínio	1
2.1. <i>Genética Molecular da Tolerância ao Alumínio em Sorgo</i>	2
3. Mapeamento Associativo	3
3.1. <i>Cálculo do Desequilíbrio de Ligação</i>	4
3.2. <i>Extensão Física do Desequilíbrio de Ligação</i>	6
3.3. <i>Populações de Mapeamento Associativo em Sorgo</i>	7
3.4. <i>Fatores que afetam a Detecção de Associação</i>	8
4. Domesticação e Dispersão do Sorgo	9
5. Objetivos	12
6. Literatura Citada.....	12
Capítulo 1	18
Population Structure for Aluminum Tolerance in Sorghum	18
1. Introduction	18
2. Material and Methods.....	21
2.1. <i>Genetic Stocks</i>	21
2.2. <i>Assessment of Al tolerance in Nutrient Solution</i>	21
2.3. <i>DNA Isolation and PCR Amplification</i>	23
2.4. <i>Data Analysis</i>	24
2.4.1. <i>Statistical analysis of Al tolerance data</i>	24
2.4.2. <i>Genetic Analysis of Al Tolerance</i>	24
2.4.3. <i>Genetic diversity analysis</i>	25
2.4.4. <i>Analysis of population structure</i>	25
2.4.5. <i>Analysis of Al Tolerance with Respect to Clusters Defined by STRUCTURE</i>	25
3. Results	26
3.1. <i>Al Tolerance Variation</i>	26
3.2. <i>Genetic Analysis of Al Tolerance</i>	28
3.3. <i>Analysis of Genetic Diversity Based on SSR Markers</i>	28
3.4. <i>Analysis of Al Tolerance with Respect to Clusters Defined at k=6</i>	30

4. Discussion	30
5. Acknowledgements	33
6. References	33
7. Tables and Figures.....	38
Capítulo 2	69
Mapeamento Associativo para Tolerância ao Alumínio em Regiões Candidatas do Loco Alt_{SB} de Sorgo	69
1. Introdução.....	69
2. Material e Métodos.....	71
2.1. <i>Material Genético</i>	71
2.2. <i>Variáveis Avaliadas</i>	72
2.3. <i>Sequenciamento e Genotipagem dos Polimorfismos na Região Alvo</i>	72
2.4. <i>Desequilíbrio de Ligação</i>	75
2.5. <i>Análise de Associação</i>	75
2.5.1. <i>Modelo Linear Misto</i>	75
2.5.2. <i>Estrutura de População e Parentesco</i>	76
2.5.3. <i>Otimização do Modelo Linear Misto para Controle do Erro Tipo I</i>	76
2.5.4. <i>Influência das Diferentes Subpopulações e Classes de Tolerância ao Al nos Resultados da Análise Associativa</i>	77
2.6. <i>Diversidade na Região Alvo</i>	77
2.7. <i>Análise Filogenética da Região Alvo</i>	78
3. Resultados	78
3.1. <i>Padrão de Desequilíbrio de Ligação na Região do Gene Alt_{SB}</i>	78
3.2. <i>Estrutura de População e Controle do Erro Tipo I</i>	81
3.3. <i>Análise de Associação</i>	84
3.4. <i>Diversidade na Região Alvo</i>	95
3.5. <i>Análise Filogenética da Região Alvo</i>	100
3.6. <i>Estratégias de Seleção</i>	101
4. Discussão.....	103
5. Anexo	108
6. Literatura Citada.....	109

RESUMO

CANIATO, Fernanda Fátima, D.Sc., Universidade Federal de Viçosa, maio de 2009.

Estrutura de população e mapeamento associativo para tolerância ao alumínio em regiões candidatas do loco Alt_{SB} de sorgo. Orientador: Aluísio Borém de Oliveira. Co-orientadores: Jurandir Vieira de Magalhães e Cláudia Teixeira Guimarães.

A toxidez causada pelo alumínio (Al) é um dos fatores que mais restringem o crescimento e o desenvolvimento das culturas em solos ácidos. A frequência de acessos de sorgo tolerantes ao Al é considerada baixa, variando entre 6 a 8% e diferenças alélicas no loco Alt_{SB} de sorgo são responsáveis por até 80% da variação fenotípica da característica, embora a presença de outros locos controlando a característica tenha sido considerada. Neste contexto, a identificação de padrões de ocorrência da característica no contexto da diversidade genética da espécie bem como a identificação dos polimorfismos associados com as variações alélicas na região do loco são importantes para a identificação e incorporação desta característica em germoplasma elite e assim constituem os objetivos deste trabalho. A análise da estrutura de população baseada em 38 locos SSR indicou que os 254 acessos de sorgo foram organizados em seis subpopulações e a análise da tolerância ao alumínio no contexto populacional apresentou forte evidência de que a tolerância ao Al esteja relacionada com a diferenciação racial, uma vez que os acessos tolerantes foram detectados em frequência superior nas subpopulações pertencentes à raça guinea. De posse dos polimorfismos candidatos dentro de uma região de 24,6kbp onde está localizado o loco Alt_{SB} , o desequilíbrio de ligação foi estimado e o mapeamento associativo foi realizado utilizando um modelo linear misto com correção adequada do Erro Tipo I. O desequilíbrio de ligação na região alvo apresentou queda para valores próximos a zero dentro de 5000pb, o que impossibilitou o isolamento das contribuições individuais dos polimorfismos causativos. No entanto, polimorfismos altamente associados com os diferentes índices fenotípicos utilizados para avaliar a tolerância ao Al foram detectados ao longo da região de 24,6kbp amostrada. Destaque foi dado aos polimorfismos localizados no segundo intron do gene Alt_{SB} , particularmente àquele localizado na coordenada 6083pb, que apresentou a contribuição mais elevada para o índice global de tolerância, representado pelo primeiro componente principal (CP1), e para o índice de indução do crescimento radicular (índice b), respondendo por 10.8 e 13.5% da variação destas características, respectivamente. Este loco discrimina 8 dos 14 acessos altamente tolerantes ao alumínio e possui uma baixa frequência no conjunto de acessos avaliado (10%). Tais características fazem desse um loco promissor

para a seleção assistida por marcadores, que pode ser combinado com outros locos para ampliar a probabilidade de amostrar acessos tolerantes em painéis de ampla diversidade, como coleções nucleares ou populações de melhoramento (i.e. mineração de alelos). O conjunto de dados apresentado fornece subsídios para futuros trabalhos com o objetivo de desvendar a estrutura funcional do loco Alt_{SB} quanto à contribuição dos polimorfismos na tolerância ao Al, além de ter contribuído para o desenvolvimento de marcadores específicos para direcionar a busca, seleção e piramidação de novas fontes de tolerância ao Al em sorgo.

ABSTRACT

CANIATO, Fernanda Fátima, D.Sc., Universidade Federal de Viçosa, May, 2009.
Population structure and association mapping for aluminum tolerance in candidate regions of the *Alt_{SB}* locus in sorghum. Advisor: Aluizio Borém de Oliveira. Co-advisers: Jurandir Vieira de Magalhães and Cláudia Teixeira Guimarães.

Aluminum (Al) toxicity is one of the most limiting factors for crop growth and development on acid soils. The frequency of Al tolerance accessions in sorghum is low, ranging from 6 to 8%, and allelic variation at the *Alt_{SB}* locus accounts for up to 80% of the phenotypic variation of the trait in some mapping populations, even though the presence of other loci controlling the tolerance has been considered. In this context, identifying patterns underlying the distribution of highly Al tolerant accessions in the context of the species genetic diversity as well identifying polymorphisms associated with allelic variation at the *Alt_{SB}* locus are the goals of this work. Population structure analysis based on 38 SSR markers indicated that the 254 sorghum accessions were organized into six subpopulations and the analysis of aluminum tolerance in this population showed strong evidence that Al tolerance is related to racial differentiation, since the tolerant accessions were detected at higher frequency in subpopulations belonging to the guinea race. With candidate polymorphisms within the 24.6kpb region where the *Alt_{SB}* gene is located, pairwise estimates of linkage disequilibrium were obtained and association mapping was performed using a linear mixed model for appropriate correction of Type I Error. Linkage disequilibrium in the target region decayed to values close to zero within 5000pb, which prevented us to isolate the individual contributions of causative polymorphisms on Al tolerance. However, polymorphisms highly associated with different phenotypic indices used to assess Al tolerance were detected along the sampled region. Emphasis was given to polymorphisms located in the second intron of the *Alt_{SB}* gene, particularly that located at the 6083pb coordinate, which had the greatest contribution to an overall tolerance index represented by the first principal component (PC1), and to an index that reflects induction of root growth upon Al exposure (b index), explaining 10.8 and 13.5% of variation for these traits, respectively. This locus discriminates 8 of the 14 accessions that are highly tolerant to aluminum and occurs at a low frequency in this panel (10%). These features make it a promising locus for marker-assisted selection, and particularly in combination with other loci, can be used to increase the probability of sampling tolerant accessions in widely diverse panels as core collections or breeding populations (i.e. allele mining). The

current dataset provides support to future work aimed at revealing the functional structure of the *Alt_{SB}* locus regarding the contribution of polymorphisms in the Al tolerance response, and is also useful to the development of specific markers to guide the identification, selection and pyramiding of new sources of Al tolerance in sorghum.

Introdução Geral

1. Importância do sorgo

O sorgo (*Sorghum bicolor* L. Moench) tem se destacado como um importante cereal, sendo cultivado para a produção de grãos e de bebidas visando o consumo humano, de biomassa para a produção de combustível e alimentação animal. Sua produção, na safra de 2007, foi de 65 milhões de toneladas em aproximadamente 43 milhões de hectares, fazendo do sorgo o quinto cereal mais produzido no mundo (FAO, 2009). O cultivo de sorgo apresenta uma grande vantagem em relação às outras culturas, pois é perfeitamente possível em áreas marginais devido a tolerância à seca apresentada pela espécie. Diante disso, esforços direcionados à maior tolerância aos fatores de estresses abióticos que limitam o seu cultivo são justificados. Dentre esses a tolerância ao alumínio (Al) tóxico do solo assume particular importância no Cerrado brasileiro.

2. Efeito da Toxidez do Alumínio

Solos que apresentam valores de pH menores do que 5,0, denominados solos ácidos, ocupam aproximadamente 50% das terras potencialmente aráveis do mundo (von Uexküll e Mutert 1995). Em valores baixos de pH, formas monoméricas parcialmente protonadas de alumínio (Al), as quais exercem efeitos fitotóxicos, são liberadas na solução do solo a partir de compostos como $\text{Al}(\text{OH})_3$ e silicatos de Al (Martin, 1992). Plantas intoxicadas por Al apresentam rápida inibição do crescimento radicular (Kochian 2004), resultando em raízes pouco desenvolvidas e incapazes de explorar camadas mais profundas de solo. Os efeitos fitotóxicos do Al resultam na restrição da absorção de nutrientes e susceptibilidade à seca (Foy et al. 1993), constituindo-se essa em uma das principais causas da baixa produtividade das culturas em solos ácidos.

É bastante claro que os efeitos tóxicos causados pelo Al são consequência da sua ação no sistema radicular das plantas (Taylor et al. 1998). Apesar de várias hipóteses acerca da tolerância ao Al terem sido formuladas ao longo de vários anos (Kochian 2004), o mecanismo fisiológico melhor embasado por dados experimentais envolve a liberação de ácidos orgânicos pelos ápices radiculares de genótipos tolerantes (Delhaize et al. 1993a, b). Os ácidos orgânicos, dentre os quais destacam-se o citrato e o malato, formam compostos

não-tóxicos estáveis com o Al^{3+} presente na rizosfera, conferindo assim tolerância à toxidez causada pelo Al (Ma et al. 2001).

2.1. Genética Molecular da Tolerância ao Alumínio em Sorgo

Em sorgo, a herança da tolerância ao Al em uma população $F_{2:3}$ derivada do cruzamento da linhagem tolerante, SC283, e sensível, BR007, foi compatível com a presença de um gene de efeito maior, que apresentou dominância parcial nas condições adotadas para avaliação do fenótipo (Magalhães et al. 2004). Os estudos de genômica comparativa relatados por esses autores indicaram que os genes *Alt_{SB}* em sorgo e *Alt_{BH}* em trigo são provavelmente frutos de mutações em locos ancestrais distintos, enquanto um QTL de efeito maior para tolerância ao Al detectado no cromossomo 1 de arroz em vários estudos (Wu et al. 2000, Nguyen et al. 2001, 2002, 2003) é possivelmente ortólogo ao gene *Alt_{SB}* em sorgo.

A herança da tolerância ao Al em sorgo foi estudada por Caniato et al (2007) por meio de análises genéticas em 12 populações F_2 com marcadores do tipo STS (*Sequence-Tagged Sites*) flanqueando o gene *Alt_{SB}*. Os autores relataram que a grande variabilidade fenotípica para tolerância ao Al observada entre as 12 linhagens de sorgo pôde ser explicada por uma série alélica no loco *Alt_{SB}* e pela presença de outros genes distintos.

Utilizando a técnica de clonagem posicional em sorgo, Magalhães et al (2007) localizaram o gene *Alt_{SB}* em um fragmento de 24,6kpb (kilo pares de bases), onde apenas três genes foram detectados. Um desses genes, que corresponde a um membro da família MATE (*Multidrug and Toxin Compound Extrusion Family*), foi expresso nos ápices radiculares de uma linhagem isogênica tolerante, mas não naqueles do par sensível, enquanto que os dois outros genes na região somente foram expressos na parte aérea de ambas as linhagens. Além disso, a expressão do gene *MATE* em *Sorghum bicolor* (denominado *SbMATE*) foi induzida pelo tratamento com Al, com aumento dos níveis de expressão ocorrendo ao longo do tempo de exposição ao Al. De maneira similar, a tolerância ao Al avaliada pela inibição do crescimento radicular em solução nutritiva também foi induzida pelo Al na linhagem tolerante, que foi seguida por um aumento na exudação de citrato pelos ápices radiculares. Essas respostas não foram observadas na linhagem isogênica sensível, que apresentou inibição drástica do crescimento radicular causada por Al, taxas de exudação de citrato extremamente reduzidas e níveis de expressão do gene *SbMATE* abaixo do limite de detecção. Dado o envolvimento de membros da família MATE no transporte de citrato e as respostas indutivas observadas para a expressão

do gene *SbMATE*, a exudação de citrato e a tolerância ao Al somente na linhagem isogênica tolerante, os dados indicaram que *SbMATE* é um transportador de citrato que corresponde ao gene *Alt_{SB}*. Finalmente, um experimento de transformação genética conduzido em *Arabidopsis* indicou que plantas transgênicas expressando *SbMATE* apresentaram maior tolerância ao Al e exudação de citrato do que os controles, fornecendo suporte experimental que um membro da família MATE em *Sorghum bicolor*, *SbMATE*, é um transportador de citrato induzido pelo Al que confere tolerância ao Al por meio do loco *Alt_{SB}*.

A sequência da região completa de 24,6kpb foi obtida para os parentais da população de mapeamento SC283 e BR007, respectivamente tolerante e sensível ao Al. Nesse trabalho, os polimorfismos identificados foram utilizados para análise de associação em um conjunto de 210 acessos de sorgo de origem diversa (Deu et al. 2006) e de 45 linhagens de programas de melhoramento do Brasil e Estados Unidos (Caniato et al. 2007).

3. Mapeamento Associativo

O mapeamento associativo explora o desequilíbrio de ligação, ou seja, a associação não aleatória de alelos entre dois ou mais locos (Lewontin e Kojima, 1964), para identificar variantes responsáveis por características complexas. Uma vez que polimorfismos causativos cosegregam com o fenótipo, alelos fisicamente próximos e em desequilíbrio de ligação causado por ligação física com os primeiros, que são também chamados de *Quantitative Trait Nucleotides* (QTNs), podem ser utilizados para localizar um gene responsável por um dado fenótipo. Muitas vezes, o mapeamento associativo pode conferir resolução superior à obtida com os mapas de ligação clássicos (Buckler e Thornsberry 2002, Flint-Garcia 2003). Para mapear QTLs são utilizadas populações derivadas de cruzamentos bi-parentais. Entretanto, o número de meioses explorado com essas populações é relativamente pequeno, fazendo com que marcadores fisicamente distantes do QTL permaneçam geneticamente associados a ele, comprometendo a individualização dos efeitos e posterior clonagem do(s) gene(s) responsáveis pelos QTLs. O uso de linhagens recombinantes endogâmicas (*Recombinant Inbred Lines* ou RILs) em F₆ ou gerações mais avançadas possibilita a redução do desequilíbrio entre marcadores ao longo das várias gerações de autofecundação. Todavia, em linhas gerais, os níveis de desequilíbrio de ligação nessas populações são ainda muito superiores àqueles observados em populações sujeitas ao acasalamento aleatório (Mackay e Powell 2007). Diferentemente, no mapeamento associativo são usadas populações naturais, coleções de germoplasma ou conjunto de materiais elite de programa de melhoramento, onde as oportunidades de

recombinações foram maiores, o que permite a localização de polimorfismos a distâncias relativamente próximas do gene de interesse. Como por exemplo, o mapeamento do loco *Vegetative to generative transition 1*, *Vgt1*, responsável pelo controle do tempo de florescimento em milho, foi localizado a uma distância 70kb de um fator de transcrição que altera a sua expressão (Salvi et al. 2007). Estas diferenças não fazem do mapeamento associativo melhor que o mapeamento de QTL e ambos devem ser considerados como métodos complementares.

3.1. Cálculo do Desequilíbrio de Ligação

O desequilíbrio de ligação pode ser estimado pelo coeficiente padronizado de desequilíbrio (D' , Lewontin 1964) e pelo quadrado do coeficiente de correlação entre frequências alélicas (r^2 , Hill e Robertson, 1968). Essas duas estimativas de desequilíbrio de ligação dependem do cálculo do coeficiente de desequilíbrio, D , que mede o desvio das frequências haplotípicas da condição de equilíbrio de Hardy-Weinberg. Para ilustrar como é obtido o D , consideremos dois locos bialélicos, A_1/A_2 e B_1/B_2 , com as frequências haplotípicas de acordo com a Tabela 1.

Tabela 1. Frequência dos quatro gametas para o caso de dois locos bialélicos.

Gametas	Frequência	Alelo	Frequência
A_1B_1	x_{11}	A_1	$p_1 = x_{11} + x_{12}$
A_1B_2	x_{12}	A_2	$p_2 = x_{21} + x_{22}$
A_2B_1	x_{21}	B_1	$q_1 = x_{11} + x_{21}$
A_2B_2	x_{22}	B_2	$q_2 = x_{12} + x_{22}$

Assumindo uma associação aleatória, a frequência de cada um dos quatro gametas será:

$$x_{11} = p_1q_1$$

$$x_{12} = p_1q_2$$

$$x_{21} = p_2q_1$$

$$x_{22} = p_2q_2$$

Quando a associação é não aleatória, deve-se acrescentar o desvio causado pela não associação dos alelos nos gametas, D (Tabela 2):

$$x_{11} = p_1q_1 + D$$

$$x_{12} = p_1q_2 - D$$

$$x_{21} = p_2q_1 - D$$

$$x_{22} = p_2q_2 + D$$

Tabela 2. Frequência dos quatro gametas para o caso de dois locos bialélicos, considerando associação não aleatória.

	A ₁	A ₂	Total
B ₁	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q ₁
B ₂	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q ₂
Total	p ₁	p ₂	

D é a diferença entre a frequência dos gametas que carregam o par de alelos dos dois locos e o produto das frequências destes alelos.

$$D = x_{11} - p_1q_1$$

Substituindo p_1 e q_1 , temos:

$$D = x_{11}x_{22} - x_{12}x_{21}$$

Com base nos valores de D , é possível calcular as duas principais medidas de desequilíbrio de ligação D' (Lewontin, 1964) e r^2 (Hill e Robertson, 1968).

O coeficiente padronizado desequilíbrio de ligação D' (Lewontin, 1964) é o valor da razão de D comparado com o seu valor máximo:

$$D' = \frac{D}{D_{\max}}$$

$D_{\max}$ é o valor máximo de D para um dado conjunto de frequências alélicas para um loco bialélico, de acordo com:

Se $D \geq 0$, $D_{\max}$ é o menor valor de p_1q_2 ou p_2q_1 ;

Se $D < 0$, $D_{\max}$ é o menor valor de p_1q_1 ou p_2q_2 .

Outra maneira de calcular o desequilíbrio de ligação é com r^2 (Hill e Robertson, 1968), que é um coeficiente de correlação que mede a associação entre os alelos dos locos A e B.

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

D' e r^2 estão envolvidos em diferentes aspectos do desequilíbrio de ligação. r^2 reflete históricos de recombinação e mutação da população, enquanto D' reflete somente o histórico de recombinação. Ambos D' e r^2 serão iguais a zero na condição de equilíbrio de ligação gerado pela recombinação entre locos (Figura 1a). D' e r^2 serão iguais a 1 quando duas mutações ligadas ocorrerem a um similar momento e não ocorrer recombinação entre elas (Figura 1b). D' será igual a 1 e r^2 menor que 1 quando mutações forem observadas em diferentes linhagens e não ocorrer recombinação entre elas, refletindo a mesma história de recombinação mas diferentes histórias mutacionais (Figura 1c).

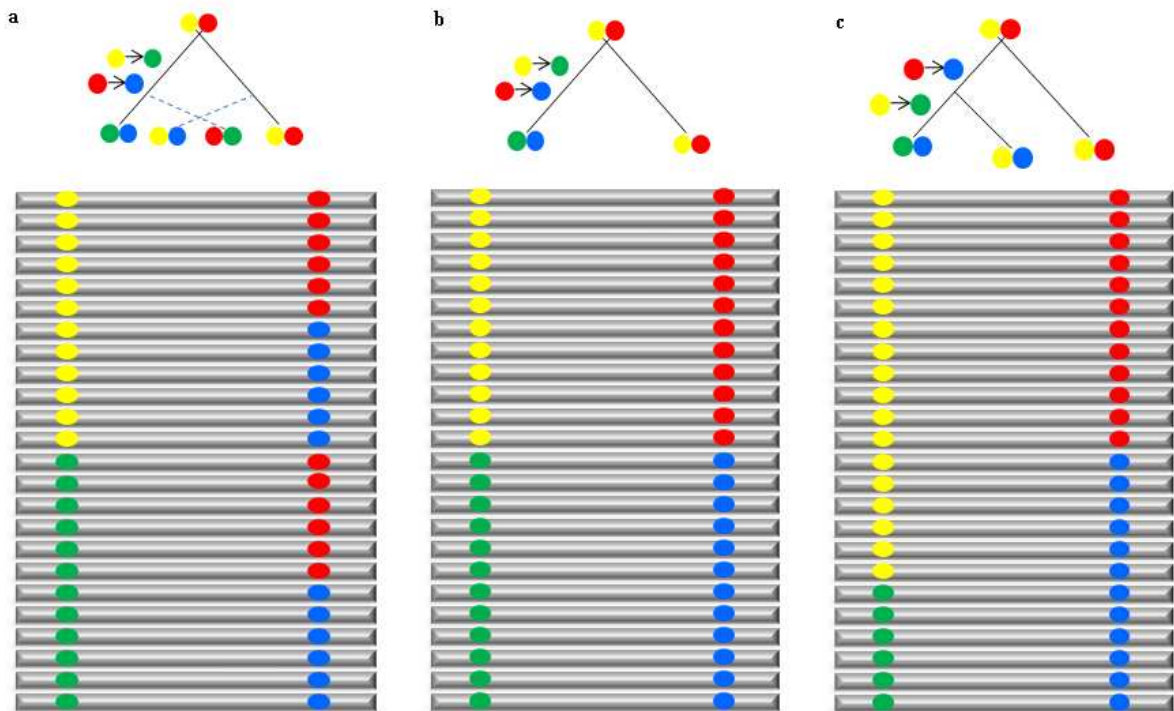


Figura 1. Hipotéticos cenários de desequilíbrio de ligação entre locos ligados causados por diferentes históricos de mutação e recombinação (Adaptado de Flint-Garcia et al. 2003)

3.2. Extensão Física do Desequilíbrio de Ligação

A extensão do desequilíbrio de ligação é um tópico que merece atenção tanto no contexto do mapeamento associativo quanto na elaboração de estratégias de seleção

assistida. Isso porque que a sua extensão define o número de marcadores necessários para a obtenção de uma cobertura genômica associada a uma alta probabilidade de identificar marcadores em desequilíbrio de ligação com genes responsáveis pelo fenótipo alvo. Para espécies de fecundação cruzada como o milho, o declínio do desequilíbrio de ligação para valores de r^2 próximos a 0,1 foi detectado dentro de 1.500 pares de bases (pb) em quatro dos seis genes estudados por Remington et al (2001) e dentro de 700pb na região do gene *dwarf8* (Thornsberry et al. 2001). Já em sorgo, o desequilíbrio de ligação avaliado em seis clones genômicos de BAC (cromossomo artificial de bactéria) indica a sua queda após 15.000pb (Hamblin et al. 2005). A manutenção do desequilíbrio de ligação é bastante variável em regiões gênicas de sorgo, para o gene *Opaco2*, *O₂*, desequilíbrio de ligação completo foi observado dentro de 4.250pb enquanto que na região do gene *Amylose extender1*, *Ae1*, desequilíbrio de ligação completo foi mantido dentro de 1.380pb (de Alencar Figueiredo et al. 2008). Em um cenário em que o desequilíbrio de ligação é mantido por pequenas distâncias como em milho, o isolamento dos polimorfismos causais é possível, como o exemplo da recente identificação de quatro polimorfismos do loco *lycopene epsilon cyclase* (*lcyE*), que juntos respondem por 58% da variação presente na via de produção de vitamina A (Harjes et al. 2008). Entretanto a cobertura genômica de espécies como o milho irá requerer grande número de marcadores. Por outro lado, quando o desequilíbrio de ligação é mantido por longas distâncias, como em sorgo, o isolamento do efeito do QTN é comprometido, uma vez que a associação detectada pode ser devido ao QTN ou a outro loco em desequilíbrio de ligação com o ele. Entretanto, a cobertura genômica demanda um número relativamente menor de marcadores em relação ao milho.

3.3. Populações de Mapeamento Associativo em Sorgo

A população de mapeamento, no contexto de uma análise por associação, deve ser representativa da diversidade genética da espécie, visando explorar eventos de recombinação históricos que irão fornecer uma resolução suficientemente elevada para detectar associações entre os marcadores e a característica de interesse. A diversidade genética em sorgo é caracterizada por padrões geográficos e raciais (de Oliveira et al. 1996, Deu et al. 1994, 2006, Barnaud et al. 2007). A análise de diversidade conduzida no conjunto de acessos de sorgo desenvolvidos pelo *Centre de Cooperation Internationale en Recherche Agronomique pour le Développement* (CIRAD - Montpellier, França) e pelo *International Crops Research Institute for the Semi-Arid Tropics* (ICRISAT - Patancheru, Índia), amostrados de acordo com a classificação racial, latitude de origem, resposta ao

comprimento do dia e forma de cultivo, mostrou ser esse painel promissor para análises de associação pela representatividade da diversidade genética da espécie (Deu et al. 2006). Recentemente, outro painel de sorgo formado por 377 acessos, com 60% de linhagens do programa de conversão de sorgo (linhagens SC) e 40% de linhagens elites voltadas para a produção de grãos e forragem e de linhagens de importância genética e histórica, mostrou-se também adequado para estudos de associação (Casa et al. 2008). Linhagens SC (i.e. sorghum converted) são caracterizadas por apresentarem porte baixo e florescimento independente do comprimento do dia, e foram desenvolvidas para permitir o cultivo de sorgo em regiões temperadas. Elas foram desenvolvidas a partir de linhagens de porte alto e sensíveis ao fotoperíodo por meio da introgressão de alelos de nanismo e insensibilidade ao fotoperíodo de um doador comum (Stephens et al. 1967). A eficiência deste último painel de associação foi recentemente verificada no mapeamento de genes que influenciam a altura de plantas em sorgo (Brown et al. 2008).

3.4. Fatores que afetam a Detecção de Associação

Estrutura de população e relacionamento familiar ou parentesco são considerados dois importantes fatores que limitam os estudos de associação. O termo estrutura de população refere-se à presença de sub-grupos com frequências alélicas diferentes dentro da população (Pritchard 2001), o que pode ser causado por fatores como domesticação e adaptação. A mistura intencional ou não intencional de indivíduos entre subpopulações que apresentam frequências alélicas diferentes cria desequilíbrio de ligação entre locos não ligados. Para visualizar este cenário, vamos pensar em uma subpopulação que é, por exemplo, fixa para os alelos A_1 e B_1 , enquanto que uma outra subpopulação é fixa para os alelos A_2 e B_2 , não ligados. Quando estas duas subpopulações que apresentam cada uma o haplótipo A_1B_1 e o haplótipo A_2B_2 entram em contato, estes haplótipos irão nos indicar completo desequilíbrio de ligação entre os locos A e B , quando de fato ele não existe. Neste sentido, para reduzir o número de falsos positivos em estudos de associação deve-se identificar e controlar estatisticamente a estrutura de população. Atualmente, é possível controlar estatisticamente a estruturação da população pelo uso de marcadores não ligados e distribuídos aleatoriamente no genoma. Esses marcadores permitem obter inferências sobre o número de sub-grupos na população, bem como a alocação de indivíduos dentro de cada sub-grupo com base na frequência alélica de cada loco marcador (Pritchard et al. 2000). Outro importante fator a ser controlado nas análises de associação é a presença de indivíduos relacionados por descendência, que compartilham alelos em comum gerando

desequilíbrio de ligação entre locos não ligados e, consequentemente, associações falsas. Recentemente, Casa et al (2008), conduzindo estudos de associação em sorgo, observaram que modelos mistos controlando a estruturação da população (Q) e o relacionamento entre indivíduos (k) reduziram bastante a incidência do Erro Tipo I, aumentando a precisão das estimativas de associação.

4. Domesticação e Dispersão do Sorgo

Acredita-se que a domesticação do sorgo tenha ocorrido no noroeste da África há pelo menos 5.000-6.000 anos atrás, a partir da espécie selvagem *Sorghum verticilliflorum* dando origem à raça mais primitiva, chamada bicolor (Kimber 2000). A diversidade apresentada entre os membros da espécie foi criada pela seleção disruptiva, por isolamento e recombinação em vários ambientes e pela seleção para diferentes finalidades. A domesticação do sorgo foi fortemente influenciada pela migração humana, pelo comércio e pelas rotas marítimas através da África e do meio-leste da Índia (Kimber 2000). A Figura 2a mostra a distribuição das raças selvagens de sorgo e na Figura 2b estão representadas as raças cultivadas na África. Comparando o padrão de distribuição das raças selvagens e cultivadas pode-se observar uma concentração de raças cultivadas entre o leste e o oeste e entre o norte e o sul, que suporta a hipótese de dispersão pela migração humana através do Sahel e do Sudão, e em direção ao sul da região do vale do Nilo, ao longo do Great Rift (Kimber 2000).

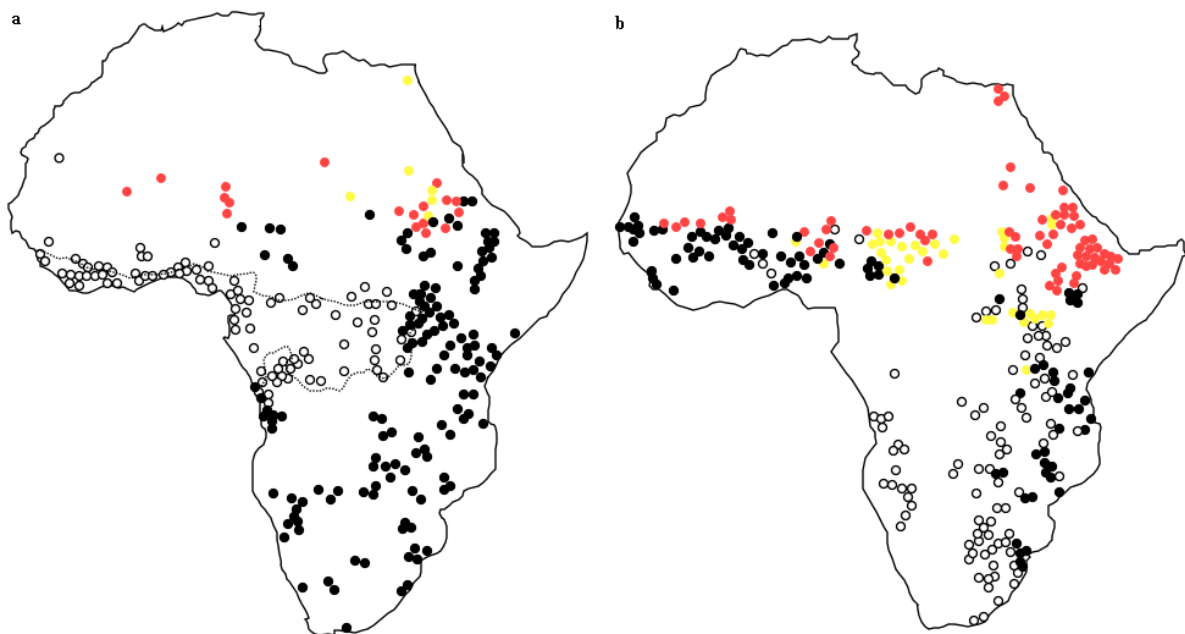


Figura 2. (a) Distribuição de variedades selvagens de *Sorghum bicolor*: var. *aethiopicum*, círculo vermelho; var. *arudinaceum*, círculo claro; var. *verticilliflorum*, círculo preto; var. *virgatum*, círculo amarelo. **(b)** Raças cultivadas: kafir, círculo claro; guinea, círculo preto ; durra, círculo vermelho; caudatum, círculo amarelo (Adaptado de de Wet e Harlan, 1971).

A migração na direção oeste culminou com a origem da raça guinea, que posteriormente foi levada para a região sul. A migração do centro de domesticação original para a região sul culminou com a origem da raça kafir. A raça caudatum, por sua vez, tem sua origem atribuída coincidente com o centro de domesticação, tendo se espalhado para o leste e o sul da África. Membros da raça durra são encontrados predominantemente no sul da Ásia e noroeste da África, mas a sua origem ainda não é clara (Dogget 1988). A raça bicolor é considerada a menos especializada e apresenta ampla distribuição (Figura 3). A Figura 4 ilustra as cinco raças básicas de sorgo cultivado: bicolor, caudatum, kafir, guinea e durra (Harlan e De Wet, 1972).

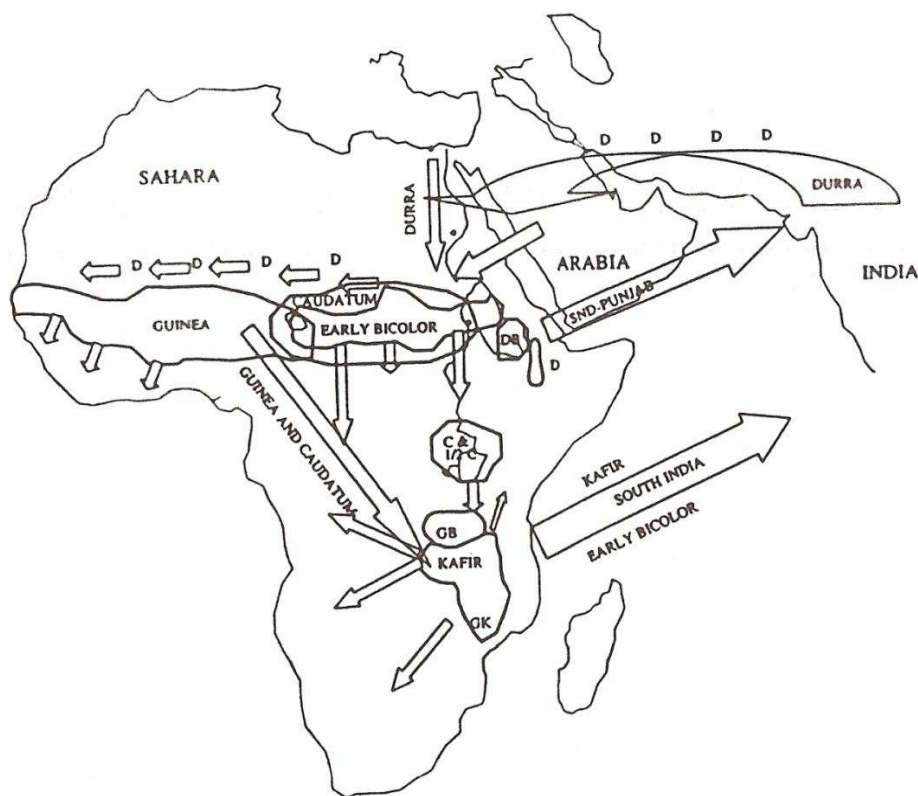


Figura 3. Áreas de origem e desenvolvimento das raças domesticadas a partir da raça bicolor primitiva e possíveis rotas de migração. C & 1/2 C (caudatum e 1/2 caudatum); DB (durra e bicolor); K (kafir); G (guinea) e GK (guinea e kafir) (Kimber 2000).

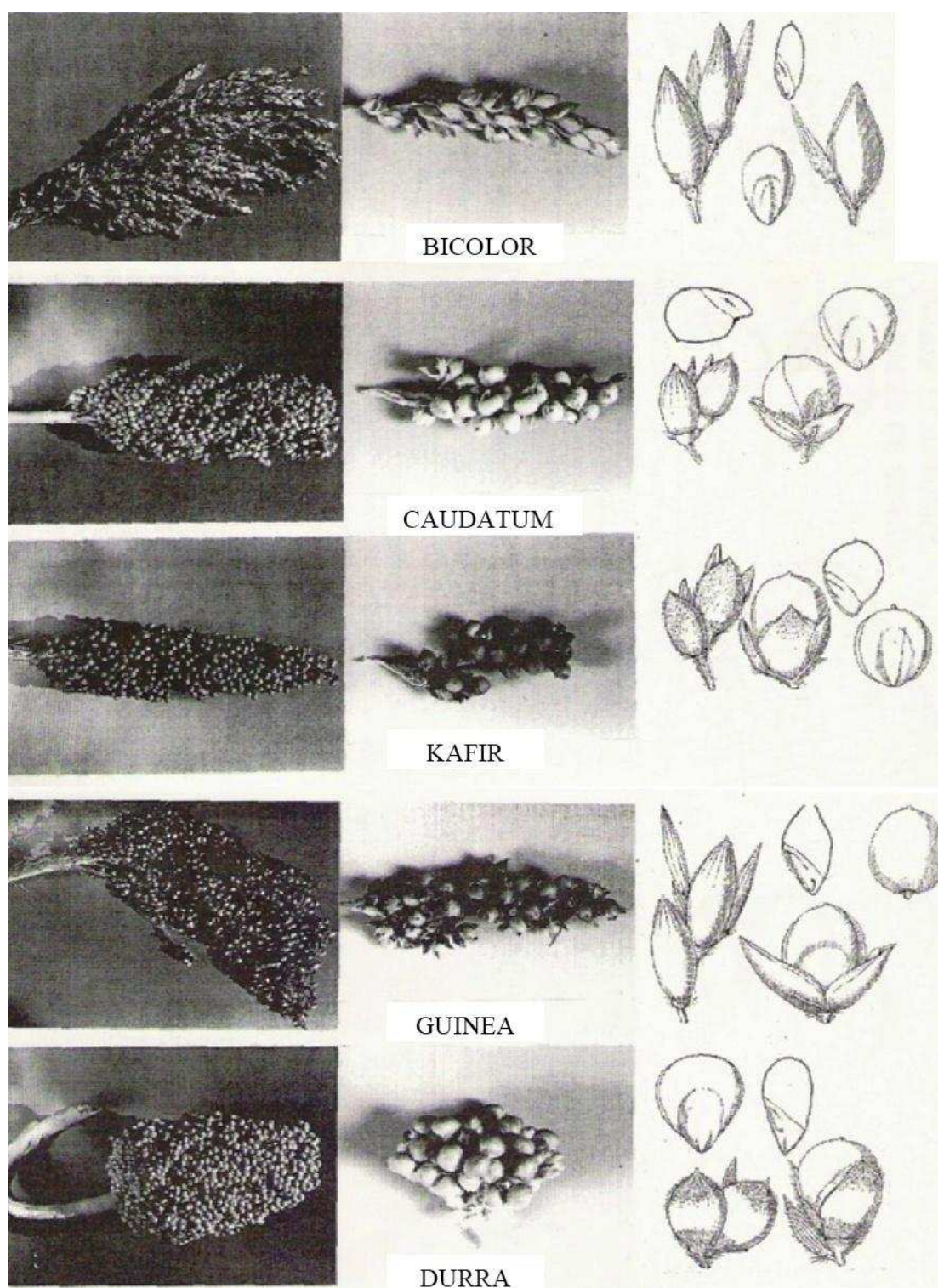


Figura 4. Ilustração das cinco raças básica de sorgo (Kimber 2000).

5. Objetivos

O presente trabalho foi realizado com o objetivo final de identificar polimorfismos funcionais no loco *Alt_{SB}*, que confere tolerância ao Al em sorgo, utilizando a estratégia de genética associativa em um painel de 254 acessos de sorgo. Assim, inicialmente foi avaliada a diversidade genética da tolerância na população alvo visando à elucidação da sua estruturação em subpopulações para uma melhor correção de associações falsas (Erro Tipo I) e a identificação de padrões de tolerância entre os acessos para construir possíveis inferências evolutivas sobre a tolerância ao alumínio em sorgo. Tais resultados estão apresentados no capítulo 1, sob a forma de um artigo científico a ser submetido para a Revista *Theoretical and Applied Genetics*.

De posse de uma coleção representativa da diversidade global da tolerância ao Al em sorgo, foi realizada uma busca pelos polimorfismos associados na região de 24,6kpb onde está localizado o gene *Alt_{SB}*, via mapeamento associativo, visando à identificação de haplótipos tolerantes e a elaboração de estratégias de seleção assistida para a piramidação da tolerância ao Al em sorgo. Os dados compõem o capítulo 2, cuja apresentação segue o mesmo formato do primeiro capítulo.

6. Literatura Citada

BARNAUD, A.; DEU, M.; GARINE, E.; MCKEY, D.; JOLY, H.I.; Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. **Theoretical and Applied Genetics**, n.114, p. 237–248, 2007.

BROWN, P.J.; ROONEY, W.L.; FRANKS, C.; KRESOVICH, S. Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. **Genetics**, v. 180, p. 629-637, 2008.

BUCKLER, E.S.; THORNSBERRY, J.M. Plant molecular diversity and applications to genomics. **Current Opinion in Plant Biology**, n.5, p. 107-111, 2002.

CANIATO, F.F.; GUIMARÃES, C.T.; SCHAFFERT, R.E.; ALVES, V.M.; KOCHIAN L.V.; BORÉM, A.; KLEIN, P.E.; MAGALHAES J.V. Genetic Diversity for Al Tolerance in Sorghum. **Theoretical and Applied Genetics**, n.114, p. 863-876, 2007.

CASA, A.M.; PRESSOIR, G.; BROWN, P.J.; MITCHELL, S.E.; ROONEY, W.L.; TUINSTRA.M.R.; FRANKS,C.D.; KRESOVICH, S. Community Resources and Strategies for Association Mapping in Sorghum. **Crop Science**, n. 48, p. 30-40, 2008.

de ALENCAR FIGUEIREDO, L.F.; CALATAYUD, C.; DUPUITS, C.; BILLOT, C.; RAMI, J.-F.; BRUNEL, D.; PERRIER, X.; COURTOIS, M.; DEU, M.; GLASZMANN, J.-C. Phylogeographic evidence of crop neodiversity in sorghum. **Genetics**, v. 179, p. 997-1008, 2008.

de OLIVEIRA, A.C.; RICHTER, T.; BENNETZEN, J.L. Regional and racial specificities in sorghum germplasm assessed with DNA markers. **Genome**, v. 39, p. 579-587, 1996.

DELHAIZE, E.; CRAIG, S.; BEATON, C.D.; BENNET, R.J.; JAGADISH, V.C.; RANDALL, P.J. I. Uptake and distribution of aluminum in root apices. **Plant Physiology**, v.103, p.685-693, 1993a.

DELHAIZE, E.; RYAN, P.R.; RANDALL, P.J. Aluminum tolerance in wheat (*Triticum aestivum* L.). II. Aluminum-stimulated excretion of malic acid from root apices. **Plant Physiology**, v.103, 695-702, 1993b.

DEU, M.; GONZALEZ-DE-LEON, D.; GLASZMANN, J.C.; DEGREMONT, I. CHANTEREAU, J.; LANAUD, C.; HAMONN, P. RFLP diversity in cultivated sorghum in relation to racial differentiation. **Theoretical and Applied Genetics**, v.88, p. 838-844, 1994.

DEU, M.; RATTUNDE, F.; CHANTEREAU, J. A global view of genetic diversity in cultivated sorghums using a core collection. **Genome**, n. 49, p.168-180, 2006

DOGGETT, H. Sorghum, 2nd ed. Wiley, New York, 1988.

FAO Statistical Database. Rome: FAO, 2009. Disponível em <<http://faostat.fao.org/faostat/collections?subset=agriculture>> Acesso em: 17 maio de 2009.

FLINT-GARCIA, S.A.; THORNSBERRY, M.J.; BUCKLER, E. Structure of linkage disequilibrium in plants. **Annual Review of Plant Biology**. v. 54, p. 357-374, 2003.

FOY, C.D.; DUNCAN, R.R.; WASKOM, R.M.; MILLER, D.R. Tolerance of sorghum genotypes to an acid, aluminum toxic tatum subsoil. **Journal of Plant Nutrition**, v.16, p.97-127, 1993.

HAMBLIN, M.T.; FERNANDEZ, M.G.S.; CASA, A.M.; MITCHELL, S.E.; PATERSON, A.H.; KRESOVICH, S. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. **Genetics**, v.171, p. 1247-1256, 2005.

HARJES, C.E.; ROCHEFORF, T.R.; BAI, L. BRUTNELL, T.P.; KANDIANIS, C.B.; SOWINSKI, S.G.; STAPLETON, A.E.; VALLABHANENI, R.; WILLIAMS, M.; WURTZEL, E.T.; YAN, J.; BUCKLER, E.S. Natural genetic variation in Lycopene epsilon cyclase tapped for maize biofortification. **Science**, v. 319, p. 330-333, 2008.

HILL, W.G.; ROBERTSON, A. Effects of inbreeding at loci with heterozygote advantage. **Genetics** v. 60, p. 615–628, 1968.

KIMBER, C.I. Origins of domesticated sorghum and its early diffusion to India and China. In: Smith CW and Frederiksen R A (eds) *Sorghum*, John Wiley & Sons, New York. pp 3–98, 2000.

KOCHIAN, L.V.; HOEKENGA, O.A.; PIÑEROS, M.A. How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. **Annual Review of Plant Biology**, v. 55, p. 459-493, 2004.

LEWONTIN, R.C.; KOJIMA, K. The evolutionary dynamics of complex polymorphisms. **Evolution**, v. 14, p. 458-472, 1960.

LEWONTIN, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. **Genetics**, v. 49, p.49–67, 1964.

MA, J.F.; RYAN, P.R.; DELHAIZE, E. Aluminum tolerance in plants and the complexing role of organic acids. **Trends Plant Sci**, v. 6, p.273–278, 2001.

MACKAY, I.; POWELL, W. Methods for linkage disequilibrium mapping in crops. Trends in **Plant Science**, v. 12, n.2, p. 57-63, 2007.

MAGALHAES, J.V.; GARVIN, D.F.; WANG, Y.; SORRELLS, M.E.; KLEIN, P.E.; SCHAFFERT, R.E.; LI, L.; KOCHIAN, L.V. Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the Poaceae. **Genetics**, v. 167, p. 1905-1914, 2004.

MAGALHAES, J.V.; LIU, J.; GUIMARÃES, C.T.; LANA, U.G.P.; ALVES, V.M.C.; WANG, Y.H.; SCHAFFERT, R.E.; HOEKENGA, O.A.; PIÑEROS, M.A. SHAFF, J.E.; KLEIN, P.E.; CARNEIRO, N.P.; COELHO, C.M.; TRICK, A.N.; KOCHIAN, L.V. A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. **Nature Genetics**, v. 39, p. 1156-1161, 2007.

MARTIN, R.B. Aluminum speciation in biology. In: DJ Chadwick, J Whelan, eds, *Aluminum in Biology and Medicine*, John Wiley and Sons, New York, pp 5-25, 1992.

NGUYEN, B.D.; BRAR, D.S.; BUI, B.C.; NGUYEN, T.V.; PHAM, L.N.; NGUYEN, H.T. Identification and mapping of the QTL for aluminum tolerance introgressed from new source, *Oryza rufipogon* Griff., into indica rice (*Oryza sativa* L.). **Theoretical and Applied Genetics**, v. 106, p. 583–593, 2003.

NGUYEN, V.T.; NGUYEN, B.D.; SARKARUNG, S.; MARTINEZ, C.; PATERSON, A.H.; NGUYEN, H.T. Mapping of genes controlling aluminum tolerance in rice: comparison of different genetic backgrounds. **Mol Genet Genomics**, v. 267, p.772–780, 2002.

PRITCHARD, J.K. STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, p. 945-959, 2000.

PRITCHARD, J.K. Deconstructing maize population structure. **Nature Genetics**. v. 28, p. 203-204, 2001.

REMINGTON, D.L.; THORNSBERRY, J.M.; MATSUOKA, Y.; WILSON, L.M., WHITT, S.R.; DOEBLEY, J.F.; KRESOVICH, S.; GOODMAN, M.M.; BUCKLER, E.S. Structure of linkage disequilibrium and phenotypic association in maize genome. **Proceedings of the National Academy of Science of USA**, v. 25, p. 11479-11484, 2001.

SALVI, S.; SPONZA, G.; MORGANTE, M.; TOMES, D.; NIU, X.; FENGLER, K.A.; MEELEY, R.; ANANIEV, E.V.; SVITASHEV, S.; BRUGGEMANN, E.; LI, B.; HAINEY, C. F.; RADOVIC, S.; ZAINA, G.; RAFALSKI, J.A.; TINGEY, S.V.; MIAO, G.H.; PHILLIPS, R.L.; TUBEROSA, R. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. **Proceedings of the National Academy of Science of USA**, v. 104, p. 11376-11381, 2007.

TAYLOR, G.J. The physiology of aluminum phytotoxicity. In: H Sigel, A Sigel, eds, Metal Ions in Biological systems: Aluminum and its Role in Biology, vol 24, Marcell Dekker, New York, p. 123-166, 1988.

THORNSBERRY, J. M.; GOODMAN, M. M.; DOEBLEY, J.; KRESOVICH, S.; NIELSEN, D.; BUCKLER, E.S. Dwarf8 polymorphisms associate with variation in flowering time. **Nature Genetics**, v.28, p.286–289, 2001.

von UEXKÜLL, H.R.; MUTERT, E. Global extent, development and economic impact of acid soils. In Plant-Soil Interactions at Low pH: Principles and Management, eds. RA Date, NJ Grundon, GE Raymet, and ME Probert. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 5-19. 1995.

WU, P.; LIAO, C.Y.; HU, B.; YI, K.K.; JIN, W.Z.; NI, J.J.; HE, C. QTLs and epistasis for aluminum tolerance in rice (*Oryza sativa* L.) at different seedling stages. **Theoretical and Applied Genetics**, v. 100, p.1295-1303, 2000.

Capítulo 1

Population Structure for Aluminum Tolerance in Sorghum

F.F. Caniato¹; C.T. Guimarães¹; R.E. Schaffert¹; L.V. Kochian²; A. Borém³; G. Avellar¹; J.V. Magalhaes*¹

¹Embrapa Maize and Sorghum, Rod. MG 424, Km 65, 35701-970, Sete Lagoas-MG, Brazil;

²U.S. Plant Soil and Nutrition Laboratory, USDA-ARS, Cornell University, Ithaca, NY

14853; ³Federal University of Viçosa, Viçosa-MG, 35570-000.

*Communicating author:

Jurandir Magalhaes

Embrapa Maize and Sorghum

Rod. MG 424, Km 65, 35701-970, Sete Lagoas-MG, Brazil

jurandir@cnpmc.embrapa.br

Phone: + 55 31 30271283

FAX: +55 31 30271279

1. Introduction

Among the various abiotic stresses that can limit sorghum production, aluminum (Al) toxicity has been deemed one of the main breeding targets on acids soil regions worldwide. This is because sorghum is intensively cultivated in these areas of the tropics and subtropics, which also show erratic rainfall. Al toxicity therein is a major constraint for yield stability and food security. When sensitive cultivars are exposed to Al in soils with pH below 5.0, ionic forms of Al are solubilized into the soil solution damaging the root system and reducing root growth, resulting in compromised water and nutrient acquisition by the plants (Kochian et al. 2004); finally resulting in severe yield losses.

The *Alt_{SB}* locus mapped in the end region of sorghum chromosome 3 was firstly identified as responsible for Al tolerance in the sorghum line, SC283, explaining 80% of the phenotypic variation in a derived mapping population (Magalhaes et al. 2004). In addition, a genetic analysis including twelve diverse sorghum lines showed that multiples alleles at the *Alt_{SB}* locus conferred a wide range of phenotypic effects for Al tolerance, although other unknown loci must play a role in sorghum Al tolerance (Caniato et al. 2007). Recently, a

gene encoding a member of the Multidrug and Toxic Compound Extrusion (MATE) family, an aluminum-activated citrate transporter, was cloned and found to underlie Al tolerance at the *Alt_{SB}* locus (Magalhaes et al. 2007).

Twenty-eight species have been described in the genus *Sorghum* (De Wet 1978) and among those, *Sorghum bicolor* ssp. *bicolor* is one of the most important staple food and fodder crop in Africa and Asia (Dogget 1988). Domestication of *Sorghum* is likely to have occurred in the northeastern part of Africa at least 5000-6000 years ago, giving rise to the early bicolor race (Kimber 2000). All cultivated sorghums belong to the bicolor subspecies and were classified into five basic botanical races, bicolor, caudatum, durra, guinea and kafir, and ten intermediates that are derivatives from intercrossing among members of the five main races, which are defined based on panicle and spikelet morphological differences (Harlan and de Wet 1972). The origin of the guinea race probably took place on wet parts of West Africa from selection acting on the original bicolor race. In fact, the typical open panicles observed among members of the guinea race are considered to be an adaptation to the more humid environment in West Africa (Kimbler 2000). From this region, the guinea race spread to eastern and southern Africa from where it was transported to Asia (de Wet et al. 1972). Guinea is the sorghum race that shows the most genetic diversity (Harlan 1975; Deu et al. 1994; 2006; de Oliveira et al. 1996; Folkertsma et al. 2005) and widest distribution (de Wet et al. 1972) and because of that it is considered to be older in relation to the other specialized races. The guinea race accounts for more than 70% of sorghum cultivated in West and Central Africa and may account for more than 50% of all sorghum produced in Africa (Folkertsma et al. 2005). The caudatum race is probably originated in the area of original domestication of the species and then spread to West and South Africa (Kimber 2000). The kafir race probably arose in southern Africa, whereas it remains uncertain whether durra sorghums originated in Africa or Asia (Kimber 2000).

Characterizing and understanding patterns of genetic diversity of the sorghum germplasm in light of the target phenotypes can be used to design efficient strategies for expanding the genetic basis of the breeding programs, thus improving the chance for sustainable breeding success. Considering that some traits may not be randomly distributed in the species diversity continuum but rather be specific of certain compartments, the evaluation of a highly diverse core collection, which acceptably represents the species diversity, may help revealing patterns that could be used to expand the genetic basis of a target trait. For example, a high frequency of common bean accessions with resistance to white mold among those originated from south american countries allowed for the

identification of 35 accessions of the active collection, among which 57% showed resistant reaction to the disease (Miklas et al. 1999). However, screening of the US Alfafa core collection for acid soil tolerance did not reveal any relationship between global diversity and acid soil tolerance (Bouton, 1996).

In sorghum, genetic diversity studies based in representative accessions indicate that geographic distribution and/or racial origin are related with the organization of genetic diversity (Aldrich et al. 1992; Tao et al. 1993; Deu et al. 1994, 2006; Cui et al. 1995; de Oliveira et al. 1996; Menkir et al. 1997; Djè et al. 2000; Barnaud et al. 2007). The first evidence that Al tolerance could be associated with specific sorghum races was obtained after screening over 6000 sorghum accessions from the world collection (Reddy et al. 2006). Many of the Al tolerant accessions from Nigeria, Uganda and Kenya were classified as caudatum or caudatum intermediates. The guinea race and guinea-bicolor intermediate also had a higher overall percentage of tolerant accessions (Reddy et al. 2006). Our previous study on genetic diversity of Al tolerance in sorghum (Caniato et al. 2007) was based on a sorghum panel that was largely composed of breeding derivatives, with a rather small representation of accessions of strict African origin or specific racial types. Therefore, the patterns of genetic diversity obtained with SSR markers reflected pedigree relationships rather than racial or geographic origins, limiting our discussions to identity by descent relationships in breeding programs. Hence, a representative sample of the global genetic diversity in the species becomes essential to uncover possible origins of highly Al tolerant accessions. For this purpose, a sorghum panel developed by the Centre de Cooperation Internationale en Recherche Agronomique pour le Développement (CIRAD - France) and the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT, Patancheru - India) was genetically merged to the breeding panel reported in Caniato et al (2007). The former has been described by Deu et al (2006) and is part of the reference sorghum set that has been recently developed within the frame of the Generation Challenge Programme (GCP, <http://www.generationcp.org/>). A genetic diversity study in the former (Deu et al. 2006) highlighted the highly diverse nature of its members whereas the later was purposely enriched for highly Al tolerance accessions within the Embrapa Maize and Sorghum breeding program (Caniato et al. 2007). The combined, 255-member panel provides appropriate coverage of all sorghum races, latitude of origin, photoperiod response and production systems, being particularly useful to study the origin and distribution of highly Al tolerant sorghums in the frame of the general patterns of genetic diversity in the species.

In this study, these 255 sorghum accessions were phenotyped for Al tolerance in

hydroponics and the results were related with racial and geographic distribution in the context of population structure inferences as revealed by the Bayesian approach implemented in the software, STRUCTURE (Pritchard et al. 2000). The results, which were analyzed in light of the known distribution of acid soils in the African continent, allowed us to conclude that Al tolerance is a rare trait that is not randomly distributed in sorghum, being rather largely specific to subpopulations enriched with members of the guinea and caudatum race.

2. Material and Methods

2.1 Genetic Stocks

Two-hundred and nine sorghum landraces that are representative of the five major sorghum morphological races, latitudes of origin, and photoperiod responses (Deu et al. 2006) were evaluated for Al tolerance in hydroponics. Additionally, 45 inbred lines that are frequently used in breeding programs in the USA and Brazil were merged into the landrace panel to form a combined 254-member panel, which was used in this study. Eighteen lines that were found to show intermediate to high levels of Al tolerance were crossed with the Al sensitive line, BR007. Single F₁ plants derived from each cross were backcrossed to BR007 to generate backcross one F₁ (BC₁F₁) populations, whose population sizes ranged from n = 47 to n=119.

2.2. Assessment of Al tolerance in Nutrient Solution

Analysis of Al tolerance in nutrient solution was performed at 0 and 148 μM Al that correspond to free Al³⁺ activities of {0} and {27} μM Al³⁺ (values inside brackets indicate respective Al³⁺ activity), estimated with the speciation software, GEOCHEM-PC (Parker et al. 1995). Twenty seven lines that showed intermediate to high levels of Al tolerance were evaluated at {0}, {39} and {60} μM Al³⁺. The highly Al tolerant line, SC566, and the Al sensitive line, BR007, that had been previously identified as such by Caniato et al (2007) were included as controls. The experiments consisted of a completely randomized design with 2 replications and 7 plants per replication. Seeds of each genotype were germinated for 4 days and seedlings were transferred to containers with nutrient solution lacking Al at pH 4.0. After 24h of acclimatation, the initial length of each seedling's root growing in control solution (*ilc*) was measured. The solution was then replaced by nutrient solution of identical

composition but containing either no Al or {27}, {39} and {60} $\mu\text{M Al}^{3+}$ supplied as $\text{AlK}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$. Final root lengths under Al treatment (*flAl*) or control solution (*flc*) were obtained after 3 and 5 days of exposure to Al. For each inbred line, percent values of net root growth inhibition (*RNRG_{3d}* and *RNRG_{5d}*) at each Al activity of {27}, {39} and {60} $\mu\text{M Al}^{3+}$ relative to the control root growth in {0} $\mu\text{M Al}^{3+}$ were estimated by dividing the net root growth under Al treatment (*flAl-ilc*) by the net root growth without Al (*flc-ilc*).

To investigate the mode of gene action, eighteen F_1 hybrids and respective backcross (BC) families were evaluated at {27} $\mu\text{M Al}^{3+}$ including the parents of each cross as controls. The experiment consisted of a completely randomized design with 3 replications and 7 plants per replication. For families segregating for Al tolerance an independent control lacking Al cannot be employed due to genetically dissimilar nature of individual plants. Thus, intrinsic root growth rates in the absence of Al were assessed during an initial 24h growth period in control solution, on an individual plant basis. Accordingly, after the 4-day germination period, seedlings were transferred to nutrient solution lacking Al and the initial length of each seedling's root in control solution (*ilc*) was measured. Final lengths in control solution (*flc*) were recorded 24h later, following replacement of the control nutrient solution with a solution of identical composition but containing {27} $\mu\text{M Al}^{3+}$. Final root lengths under Al treatment (*flAl*) were obtained after 5 days of exposure to Al. The degree of root growth inhibition caused by Al over 5-day exposure period relative to the control root growth was calculated as $\text{RRG (\% relative root growth)} = [(flAl - flc)_{5d} / (flc - ilc)_{1d} \times 5] \times 100$ where d is the period measured in days.

Sorghum lines were inspected for root damage after 5 days Al exposure and a visual root damage (VRD) scale ranging from 1 to 5 was applied as follows: 1) absence of secondary roots, main root brown, root apex translucent, slightly bent and heavily damaged; 2) absence of secondary roots, main root showing dark spots, root apex slightly brownish, swollen and damaged; 3) absence of secondary roots, main root with occasional dark spots, root apex white, slightly swollen and slightly damaged; 4) absence of secondary roots, main root white, root apex undamaged; 5) presence of profuse secondary root growth, main root white, root apex undamaged. Three independent evaluations were carried out for estimating VRD means.

To gain insights into the pattern of root growth between 3 and 5 days of Al exposure, a linear regression model was adjusted between total root growth (mm) at each time point and the two time periods. The regression coefficients obtained from the linear regression analysis was used as an induction index (b index) where $b=0$ indicates that root growth

stalled after 3 days of Al exposure and $b > 0$ indicates induction of root growth caused by Al exposure.

2.3. DNA Isolation and PCR Amplification

Leaf tissue from three plants of each of the 45 accessions from Embrapa collection and leaf from individual seedlings of the each segregating family were used for DNA isolation according to Saghai-Maroo et al (1984). The markers CTG29 (CTG29F: HEX-ATGCAGTATCTGCAGTATCATTT; CTG29R: AATCCGTCAGGTCAGCAATC), M181 (M181F: 6FAM-AAGGCAACAACCTGAGGCACT; M181R: TCGCTAGAGTGGTGCAAGAA), S17 (S17F: GGCTGCCCCGTCCCTTTCTCTGTCT; S17R: CCGGGGCGCTGGGCTTCCTT) and S73 (S73F: AAGCGCTGGCCCAAATGAAATGA; S73R: GAGCCAACACGGGGAGAACAAGTC) were used to determine whether Al tolerance in the tolerant sources was due to allelic variation at *Alt_{SB}*. CTG29 and M181 are Sequence Tagged Site (STS) markers, which flank *Alt_{SB}* at genetic distances of 0.5 and 0.1 cM, respectively (estimate obtained in 2085 F₂ individuals derived from a cross between SC283 and BR007, Magalhaes et al. 2007). Upon positional cloning of *Alt_{SB}* (Magalhaes et al. 2007), another two markers, S17 and S73 were developed for sequences in the same bacterial chromosome (BAC) that harbors *Alt_{SB}*. S17 is located at 32.1kbp from *Alt_{SB}* in the same side as CTG29 whereas S73 is located at 22.4kbp from *Alt_{SB}* in the same side as M181.

PCR reactions with CTG29 and M181 were performed as described in Caniato et al (2007). For S17 and S73, amplifications were carried out in a reaction volume of 20 μ L which contained 30ng of genomic DNA, 2X polymerase chain reaction buffer, 2.0mM dNTP, 30mM MgCl₂, 4.0 pmol of each primer, 5% of Dimethyl sulfoxide (DMSO) and 0.5U of Taq polymerase (Phoneutria, Belo Horizonte, MG). Amplification proceeded with an initial denaturation step of 95°C for 1 min followed by 30 cycles at 94°C for 1 min, 62°C for 1 min, 72°C for 1 min, and a final extension step at 72°C for 10 min. Electrophoresis was carried out in 1% (w/v) agarose gel at 100.0 V in 1X TAE buffer.

Thirty-eight SSR markers from the sorghum SSR kit (http://sat.cirad.fr/sat/sorghum_SSR_kit), which are evenly distributed across the sorghum genome, were used for genetic diversity analysis. Fragment sizes obtained for 209 CIRAD lines were provided by the GCP and the 45 lines from the Embrapa collection were genotyped using the same SSR markers. Because differences in allele sizes for the same alleles are expected between labs, the GCP also defined a set of 10 highly diverse sorghum

lines, for which a wide range of allele sizes is available, to be used as controls. These ten lines were used for PCR amplification with the same SSR markers along with the 45 lines from the Embrapa collection to normalize differences in fragment sizes considering what was obtained in the GCP and Embrapa conditions. PCR reactions were carried out as described for STS amplifications but omitting DMSO and using 2.5 pmol of each SSR primer. Amplification proceeded with a touchdown protocol including an initial denaturation step at 94°C for 4 min followed by 9 cycles at 94°C for 45s, 60°C for 1min with a reduction rate of 0.5°C per additional cycle, 72°C for 1min 15s, and 24 cycles at 94°C for 45s, 55°C for 1min, 72°C for 1 min and 15s with a final extension step of 5 min at 72°C. Three micro liters of 200-fold diluted amplification products and 6.9 µl of Hi-Di formamide (Applied Biosystems, Foster City, CA) were mixed with 0.1µl GS500 ROX internal size standard and denatured at 95°C for 5 min. The fragments were assayed on the ABI3100 sequencer (Applied Biosystems, Foster City, CA). Fragment sizes were determined based on migration relative to the internal size standard using the GENEMAPPER 3.5 software. The allele sizes obtained from each control line were compared to the expected allele sizes posted on http://sat.cirad.fr/sat/sorghum_SSR_kit and a correction factor for each marker was imposed to normalize allele sizes for the GCP and Embrapa datasets.

2.4. Data Analysis

2.4.1. Statistical analysis of Al tolerance data

One-way analysis of variance (ANOVA) of *RNRG*_{3d}, *RNRG*_{5d}, VRD and b index data at each Al activity followed by the Scott–Knott test (Scott and Knott 1974) were undertaken to cluster the accessions into homogeneous groups. *RNRG*_{3d}, *RNRG*_{5d}, VRD and b index data at {27} µM Al³⁺ were also subjected to Principal Component Analysis (PCA, Hottelling 1933) based on standardized variables to elucidate the contribution of each variable to the global Al tolerance response.

2.4.2. Genetic Analysis of Al Tolerance

For BC families, F₁ hybrids and respective parents, statistical significance for RRG means was tested by the Student's t test. Individual χ^2 tests for goodness-of-fit to an expected 1:1 segregation ratio on BC families were performed for the markers CTG29, M181, S17 and S73. Marker-trait associations were analyzed by single-factor ANOVA of the RRG values using the two marker genotypic classes as a classification variable.

2.4.3. Genetic diversity analysis

Number of alleles, polymorphism information content (PIC) values and fitness to the Hardy Weinberg Equilibrium (HWE) were calculated among 254 sorghum accession with PowerMarker version 3.25 (Liu and Muse 2005). A Neighbor-Joining tree (Saitou and Nei 1987) was obtained from the dissimilarity matrix $d_S = 1 - S_D$ (Nei and Li 1979), which is based on the Dice similarity coefficients (S_D , Dice 1945) calculated among all pairs of accessions as determined by $SD = \frac{2v_{ij}}{2v_{ij} + w_{ij} + x_{ij}}$, where v_{ij} is the number of bands shared between both accessions, w_{ij} is the number of bands present in accession i and absent in accession j and x_{ij} is the number of bands absent in accession i and present in accession j .

2.4.4. Analysis of population structure

Bayesian cluster analysis, as implemented in the software STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), was used to estimate the number subpopulations (k) based on the SSR data set. The program accounts for Hardy-Weinberg or linkage disequilibrium in the global populations by introducing population structure (i.e. clustering individuals into subpopulations), attempting to find a grouping configuration that minimize deviations from Hardy-Weinberg as well as linkage equilibrium within subpopulations (Pritchard et al. 2000). The admixture model with correlated allele frequencies was adopted, with burn-in length 100,000 and 1,000,000 run length with five independent runs for each number of inferred groups (k), with initial k values ranging from 1 to 13. It has been reported that in some cases the log probability of data do not provide a correct estimate of number of clusters (Evanno et al. 2005; Wang et al. 2008). Thus, we calculated the second order change of log of probability of data (Δk) divided by its standard deviation, called Δk (Evanno et al. 2005). The rate of change in Δk between successive k values was adopted to identify the most likely k .

2.4.5. Analysis of Al Tolerance with Respect to Clusters Defined by STRUCTURE

Due to the unbalanced nature of the resulting subpopulations, the non-parametric Kruskal–Wallis test was used to determine whether subpopulations showed differentiation for $RNRG_{3d}$, $RNRG_{5d}$, VRD and b index at $\{27\} \mu M Al^{3+}$. Statistical significance for all pairwise differences among subpopulations to each variable was estimated by calculating

the least significant difference (lsd) between subpopulations as

$$lsd = \frac{Z\alpha}{[k(k-1)] \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}, \text{ where } \frac{Z\alpha}{[k(k-1)]} \text{ is the superior limit of the normal distribution; } n_i \text{ and } n_j \text{ are the number of individual within each subpopulation, } i \text{ and } j, \text{ respectively, and } N \text{ is the total number of individuals.}$$

3. Results

3.1. Al Tolerance Variation

Means and standard deviations for 254 sorghum lines for $RNRG_{3d}$, $RNRG_{5d}$, VRD and the b index variables are reported in Supplementary Table 1. Extensive phenotypic variation was observed for all variables. Sixteen lines or 6% of the whole panel showed $RNRG > 80\%$, thus being classified as highly Al tolerant. If we compute lines that are breeding derivatives (e.g. the sorghum line 9929034 is derived from SC566 and CMSXS226R and CMSXS227R are derived of SC283) this number is reduced to thirteen tolerant lines or 5% of the whole panel. The line IS14351 stood out both in terms of $RNRG_{3d}$ and $RNRG_{5d}$, which were estimated as 234 and 262%, respectively. For VRD, IS14351 grouped with other nineteen accessions with mean score ranging from 4 to 5. Together, IS26554 and SC175 formed a homogeneous group that showed the highest b index and IS14351, CMSXS227 (an SC283 derivative) and IS26457 formed the second group in terms of b index with the Al tolerant standard, SC566.

Accessions that showed intermediate to high Al tolerance were subsequently screened at {39} and {60} μM Al (Supplementary Table 2), which discriminated seven and four groups for $RNRG_{3d}$ and $RNRG_{5d}$, respectively. The sorghum accession IS14351 did not group with any other accessions except for the b index at {39} μM Al, again emphasizing its outstanding levels of Al tolerance. At {60} μM Al, four homogeneous groups were identified for $RNRG_{3d}$ and three for both $RNRG_{5d}$ and b index, in contrast with seven and four groups, respectively, at {39} μM Al.

Principal Component Analysis (PCA) with $RNRG_{3d}$, $RNRG_{5d}$, VRD and the b index allowed us to identify the first and second principal components as responsible for 95.7% of the total Al tolerance variation, with the first component explaining the vast majority, 82.8% of the variation (Table 1). Thus, as only marginal increases in the portion of the variance explained were observed by adding other PCs, two PCs were deemed adequate for

dimensionality reduction with satisfactory representation of the experimental space. The first principal component (PC1) can be interpreted as a general Al tolerance index, with a slightly higher importance for $RNRG_{5d}$, whereas the second principal component (PC2), explaining 12.9% of the variation, contrasts relative root growth to the induction and visual damage responses. Thus, accession scores close to zero for PC1 indicate extreme Al sensitivity whereas Al tolerance increased as PC1 scores increased. In addition, scores close to zero for PC2 indicate equilibrium between relative root growth and induction/visual damage responses whereas positive scores indicate predominance of the former over root growth responses (Supplementary Figure 1). A dense group formed by Al sensitive accessions was observed at low scores for both, PC1 and PC2. As the score for PC1 increased, highly and intermediate Al tolerant accessions (scores for PC1 > 2) showed significant dispersion for PC2 scores. For example, IS14351, was an outlier in Supplementary Figure 1, confirming its outstanding Al tolerance but its highly negative score for PC2 suggested that induction was not particularly important for this accession at {27} μ M Al. Other sorghum accessions showed intermediate to high scores for PC1 but also high scores for PC2 suggesting that Al tolerance as measured by relative root growth and the induction/visual damage responses may have at least partially independent control.

The sorghum accessions that had been previously classified for Al tolerance and racial origin were then geographically represented on the map of Africa based on latitude and longitude of origin and Figure 1 integrates all these information. When the complete dataset was analyzed, 80% of the sorghum accessions were sensitive to Al ($RNRG_{5d}$ <30%), 14% were intermediate (30%< $RNRG_{5d}$ <80%) and only 6% or sixteen sorghum accessions showed high Al tolerance ($RNRG_{5d}$ >80%). Among the 39 intermediate accessions, 27 belonged to the guinea race and 4 to the caudatum race, with 2 being intermediates between these two races and 6 belonging to other sorghum races. Among the 16 tolerant accessions, 7 belonged to the guinea race and 2 to the caudatum race, with 1 belonging to other sorghum race and 6 without racial classification (Figure 1 and Supplementary Table 1). Therefore, the majority of the Al tolerant and intermediate accessions were found among members of the guinea race. The caudatum race, although with a lower frequency of Al tolerant and intermediate accessions than the guineas, is also relevant as a repository of Al tolerant accessions. This is also reinforced by the presence of highly Al tolerant accessions such as SC566 in the caudatum race. Al tolerant and intermediate accessions were prevalent in West Africa and southeastern Africa mostly in areas with low Al saturation (0 - 25%).

3.2. Genetic Analysis of Al Tolerance

Markers flanking *Alt_{SB}* were tested for goodness-of-fit to a 1:1 segregation ratio and revealed segregation distortion for BC families having as donors IS10801, IS15443, IS21662, IS24139, IS26554 and IS27490 (Table 2).

Single marker analyses with markers flanking *Alt_{SB}* revealed that Al tolerance showed by IS10801, IS14351, IS21519, IS23645, IS23666, IS24139 and IS26554 can at least in part be attributed to the *Alt_{SB}* locus, whereas significant marker-trait association was not found for BC families derived from IS15148, IS15443, IS16396, IS21662, IS23142, IS23644, IS26457, IS27490, IS29691, SSM215 and SSM1057. However, a dramatic reduction in Al tolerance as measured by %RRG was observed when then F₁ hybrids were compared with the respective Al tolerant parents. In fact, in some cases the mean for the F₁ hybrids did not differ from that for BR007 (for example, F₁ hybrid BR007 x IS10801, Table 2), which can compromise the linkage analysis in BC families.

3.3. Analysis of Genetic Diversity Based on SSR Markers

The use of ten control lines derived from the sorghum SSR kit allowed us to join the 209 sorghum accessions provided by the GCP with the 45 accessions from the Embrapa collection. In some cases, control lines amplified in our conditions displayed the same fragment sizes to those posted on http://sat.cirad.fr/sat/sorghum_SSR_kit (see marker gpsb067, Supplementary Table 3). In other cases, a constant difference within loci was observed allowing us to normalize the fragment sizes in the two datasets. This was the case for the fragment sizes obtained with markers mSbCIR246 and Xtxp114, for which the fragments obtained in the Embrapa collection were corrected by +2 and -19bp, respectively.

A total of 501 alleles of which 399 showed frequency inferior to 10%, were revealed by 38 SSR loci across 254 sorghum accessions. The average of number of alleles per locus was 13.2, ranging from 2 for marker Xtxp136 to 29 for Xgap206. The Polymorphic Index Content (PIC) value over the 38 SSR markers averaged 0.65, ranging from 0.19 for marker mSbCIR246 to 0.93 for markers Xgap206 and Xtxp321 (Supplementary Table 4).

The clusters identified with the genetic diversity analysis were consistent with both racial and geographical origins as well as known pedigree relationships. Supplementary Figure 2 shows that the main clusters corresponded to guinea margaritifera, guinea accessions from western Africa, transplanted sorghum from Lake Chad and Cameron, SC283 and derivatives, kafir accessions from southern Africa, lines from Brazil and USA, caudatum accessions from Africa and sorghum accessions from the Embrapa collection,

caudatum and intermediates from Africa and derivatives from Brazil, caudatum from western Africa, guinea accessions from southern Africa and Asia, durra from Africa, durra accessions from Asia, caudatum and intermediate accessions from China and lines from USA. Inclusion of accessions from Embrapa collection did not substantially alter the clusters previously identified by Deu et al (2006).

The $\ln(k)$ vs k curve obtained with STRUCTURE did not appear to have clearly reached saturation when k ranged from 1 to 13, raising questions to the optimum number of subpopulations in this panel as defined with this criterion. A plateau appeared to have been reached between $k=6$ and $k=10$ with drop of $\ln(k)$ values at $k=7$ and a slight increase tendency was observed afterwards (Figure 2). The variation of $\ln(k)$ values at $k=7$ in comparison to others points suggests convergence problems with the Markov chain at $k=7$. Hence, an additional criterion based on the magnitude change of $\ln(k)$ relative to the standard deviation between successive k values (Δk , Evanno et al. 2005) was adopted. Evident breaks in the slope of the Δk distribution indicate optimum k values. Using the Δk criterion, the most evident level of differentiation was observed with $k=4$, but additional peaks, although much less evident, were also detected at $k=6$, $k=8$ and $k=12$ (Figure 3). The corresponding clusters for $k=4$ (Figure 4a) were comprised of guinea accessions from western Africa and guinea margaritifera ($k4Q1$), durra, bicolor and caudatum accessions from eastern Africa and Asia ($k4Q2$), caudatum accessions from Africa, group of transplanted caudatum and durra accessions from Lake Chad region and lines from Brazil and USA ($k4Q3$), and kafir and Guinea accessions from southern Africa ($k4Q4$). Two clusters that were defined at $k4$, $k4Q3$ and $k4Q4$, were split at $k=6$. At $k=6$, the former $k4Q3$ group was separated into $k6Q2$, caudatum accessions from Africa and group of transplanted caudatum and durra accessions from Lake Chad region and $k6Q3$, lines from Brazil and USA. In addition, $k6Q4$, kafir accessions from southern Africa were also separated from $k6Q6$, guinea accessions from southern Africa (Figure 4b). From $k=6$ to $k=8$ level, the cluster $k8Q5$, group of transplanted caudatum and durra accessions from Lake Chad region separated from $k8Q8$, caudatum accessions from Africa (Figure 4c). Finally, three other subdivisions were observed from level $k=8$ to $k=12$, with separation of $k12Q9$, (guinea margaritifera) from $k12Q11$, (guinea accessions from western Africa), $k12Q3$ (bicolor accessions) from $k12Q7$ (durra accessions from eastern Africa and Asia and caudatum accessions from China) and $k12Q12$ (some caudatum accessions from eastern Africa) from $k12Q8$ (caudatum accessions from Africa, Figure 4d).

One of the assumptions of the Bayesian algorithm implemented in STRUCTURE is the definition of clusters that, as much as possible, are not in disequilibrium and Pritchard et al (2000) argues that departure from equilibrium may result in overestimating the optimum number of clusters. Although deviations from HWE were detected for all SSR loci evaluated at each subpopulation level (data not shown), the resulting clusters did show clear biological meaning at all differentiation levels. Nonetheless, the largest proportion of individuals assigned to a specific cluster with a cluster membership probability higher than 80% was obtained for $k=4$ and $k=6$, with 81 and 71%, respectively. For both $k=8$ and $k=12$, the proportion of individuals with a cluster membership probability higher than 80% was close to 60%. Hence, we choose $k=6$ as most probable number of subpopulation by its clustering efficiency and by the fact that the two additional groups obtained in comparison to $k=4$ had clear biological consistency. However, cluster membership should not be adopted as the only criterion to define k , once hybridization between sorghum races is very common, resulting in the occurrence of intermediate races (Harlan and de Wet 1972).

3.4. Analysis of Al Tolerance with Respect to Clusters Defined at $k=6$

The Kruskal–Wallis test is described as a non-parametric ANOVA and it was used here to test if there are differences between the six subpopulations based on ranking the pooled data from $RNRG_{3d}$, $RNRG_{5d}$, VRD and b index. The results indicated that subpopulations significantly differed for all Al tolerance variables (data not shown).

The non-parametric lsd test indicated that $RNRG_{3d}$ was significantly higher for Q6, guinea accessions from southern Africa and Asia and to Q3, lines from Brazil and USA (Table 3). For $RNRG_{5d}$, VRD and b index the most differentiated populations were Q1, guinea accessions from western Africa and guinea margaritifera accessions; Q3, lines from Brazil and USA and Q6, guinea accessions from southern Africa and Asia (Table 3). The superior Al tolerance for subpopulations whose members largely belong to the guinea race agrees with the fact that 46 out of 55 guinea accessions showed intermediate to high Al tolerance (Supplementary Table 1). However the high level of tolerance showed by subpopulations with predominance of lines from Brazil and USA reflects the fact that several of those lines have been purposely selected for breeding Al tolerant sorghums (Caniato et al. 2007).

4. Discussion

Sorghum is a highly diverse species which is reflected in diverse morphological

racess, geographic origins and selection for various purposes such as grain and drinks for human consumption, biomass production for fuel, feed and fiber. The sorghum panel used in this study was assembled to represent the widest possible range of the species diversity (Deu et al. 2006) and lines that are currently used for sorghum breeding in the US and Brazil, with particular emphasis to sources of AI tolerance in the Brazilian program (Caniato et al. 2007). When genetic diversity analysis was carried out in the combined panel, clustering patterns consistent with racial and geographic origins could be observed. Many studies have indicated that the organization of sorghum diversity is linked to geographic and/or racial classifications (Aldrich et al. 1992; Tao et al. 1993; Deu et al. 1994; 2006; Cui et al. 1995; de Oliveira et al. 1996; Menkir et al. 1997; Djè et al. 2000; Barnaud et al. 2007). The main clusters observed by Deu et al (2006) with 210 accessions belonging to the world collection, which were common to our study, were not significantly altered by adding the Embrapa and US lines and the main clusters obtained here were consistent with the clusters obtained with the model based algorithm implemented in STRUCTURE. Lines from USA and from the Embrapa collection formed clusters mostly with members of the caudatum race. This is not surprising once the caudatum race has been considered the most important source of germplasm for sorghum breeding programs worldwide (Harlan and de Wet 1972).

In this study, the vast majority of the sorghum accessions were highly AI sensitive, whereas 14 and 6% of the lines showed intermediate ($RNRG_{5d}$ between 30 and 80%) and high levels of AI tolerance ($RNRG > 80\%$). This low rate of AI tolerance accessions was not surprising, since a large scale screening encompassing 6000 sorghum accessions from the world collection identified only 8% of tolerant accessions (Reddy et al. 2006). Therefore, as observed in wheat, AI tolerance in sorghum is likely a derived trait and not an inherent state in the species (Garvin and Carver 2003).

Recent studies showed that Alt_{SB} is the main locus controlling AI tolerance in sorghum, since 9 out of 11 lines evaluated by Caniato et al (2007) relied on Alt_{SB} to achieve AI tolerance. However, marker trait associations in this study were not significant for 11 lines. The lack of linkage with markers flanking Alt_{SB} should be interpreted with caution. Magalhaes et al (2004) reported a partial dominant gene action for Alt_{SB} in a population derived from the AI tolerant standard, SC283 and the AI sensitive line, BR007. In this study we observed, that the F_1 hybrids derived from the various tolerant sources and BR007 showed a dramatic reduction in AI tolerance with respect to the respective tolerance donors, remaining only slightly more tolerant than the AI sensitive parent, BR007. Thus, AI tolerance conferred by the exotic alleles coming from the world collection behaved more

like a recessive trait. This limits our conclusions with regards to genetic linkage as the heterozygous and homozygous sensitive classes are the only ones that are informative in a backcross population. Near-isogenic lines (NILs) are being developed to confirm the presence of functional *Alt_{SB}* alleles in the various tolerant sources.

There is strong evidence that Al tolerance is related with racial differentiation based on the superior phenotypic performance observed in sorghum accessions belonging to clusters formed by the guinea race, which were also associated with a higher frequency of Al tolerant lines with functional *Alt_{SB}* alleles. We observed eight members of the guinea race, one intermediate guinea/caudatum and only SC566 (caudatum, Caniato et al. 2007), carrying functional *Alt_{SB}* alleles. The much lower frequency of lines harboring functional *Alt_{SB}* alleles in caudatum in relation to guinea members suggests that Al tolerance might have been originated within the guinea being subsequently transferred to caudatum sorghums by hybridization. In fact, it is known that guinea race is sympatric with all four other basic races of sorghum and interracial hybrids among them are occasionally observed (de Wet et al. 1972). Guinea-caudatum are also commonly encountered in the drier parts from Nigeria to Uganda, suggesting that Al tolerance observed among caudatum accessions could have been transferred by hybridization. Interestingly, the high Al tolerant line, SC566, classified as caudatum clustered with guinea lines in this study, reinforcing the possibility of interracial transfer of Al tolerance genes in sorghum.

The guinea race has its origin in humid zones of western Africa, from where it spread to eastern and southern Africa and subsequently Asia (Dogget 1988). Among sorghums from the guinea race, three distinct groups are frequently observed in genetic diversity studies, guinea from West Africa, guinea from southern Africa and a third group composed of accessions with smaller seeds that were classified by Snowden (1936) as guinea *margaritiferum*. The present geographical distribution of guinea sorghums reflects the domestication history and movement of this race as guinea sorghums are believed to be originated in wet parts of west Africa from early bicolor types, with later spread to southern Africa (Kimber 2000)

By providing representative samples from the species germplasm, core collections help bridging important traits such as Al tolerance from the vast number of accessions in the world collections into breeding programs. Our results should help designing more efficient molecular strategies to expanding the genetic basis of Al tolerance in breeding programs, focusing on compartments of the sorghum genetic diversity that are enriched for this trait.

5. Acknowledgements

This work is part of the PhD thesis of the first author. We thank FAPEMIG, CNPq and CAPES for providing financial resources for her studies as well as the Generation Challenge Programme and The McKnight Foundation for financial support.

6. References

Aldrich PR, Doebley J, Schertz KF, Stec A (1992) Patterns of allozyme variation in cultivated and wild sorghum bicolor. *Theor Appl Genet* 114:451-460

Barnaud A, Deu M, Garine E, Mckey D, Joly HI (2007) Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. *Theor Appl Genet* 114:237-248

Bouton JH (1996) Screening the Alfafa core collection for acid soil tolerance. *Crop Sci* 36:198-200

Caniato FF, Guimaraes CT, Schaffert RE, Alves VMC, Kochian LV, Borem A, Klein PE, Magalhaes JV (2007) Genetic diversity for aluminum tolerance in sorghum. *Theor Appl Genet* 85:863-876

Cui YX, Xu GW, Magill CW, Schertz KF (1995) RFLP-based assay of *Sorghum bicolor* (L) Moench genetic diversity. *Theor Appl Genet* 90:787-796

de Oliveira AC, Richter T, Bennetzen JL (1996) Regional and racial specificities in sorghum germplasm assessed with DNA markers. *Genome* 39:579-587

de Wet MJM, Harlan JR (1971) The origin and domestication of *Sorghum bicolor*. *Economic Botany* 25:128-135

de Wet MJM, Harlan JR, Kurmarohita B (1972) Origin and evolution of Guinea sorghums. *East Afr Agric For J* 114-119

de Wet MJM (1978) Systematics and Evolution of *Sorghum* Sect. *Sorghum* (Gramineae).

Amer J Bot 65:477-484

Djè Y, Heuertz M, Lefèbvre C, Vekemans X (2000) Assessment of genetic diversity within and among germoplasm accessions in cultivated sorghum using microsatellite markers. *Theor Appl Genet* 100:918-925

Deu M, Gonzalez-de-Leon D, Glaszmann J-C, Degremont I, Chantereau J, Lanaud C, Hamon P (1994) RFLP diversity in cultivated sorghum in relation to racial differentiation. *Theor Appl Genet* 88:838-844

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. *Genome* 49:168-180

Doggett H (1988) *Sorghum*, 2nd edn. Wiley, New York

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611-2620

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567-1587

Folkertsma RT, Rattunde HFW, Chandra S, Raju GS, Hash CT (2005) The pattern of genetic diversity of guinea-race *Sorghum bicolor* (L) Moench landraces as revealed with SSR markers. *Theor Appl Genet* 111:399-844

Garvin DF, BF Carver (2003) Role of the genotype in tolerance to acidity and aluminum toxicity. In: Rengel Z (ed) *Handbook of soil acidity*. Marcel Dekker Inc, New York, pp 387-406

Harlan JR, de Wet MJ (1972) A simplified classification of cultivated sorghum. *Crop Sci* 12:172-176

Harlan JR (1975) Geographic patterns of variation in some cultivated plants. *The Journal of*

Heredity 66:184-191

Hotteling H (1936) Relations between two sets of variates. *Biometrika* 28:321-377.

Kimber CI (2000) Origins of domesticated sorghum and its early diffusion to India and China. In: Smith CW and Frederiksen R A (eds) *Sorghum*, John Wiley & Sons, New York. pp 3-98

Kochian LV, Hoekenga OA, Piñeros MA (2004) How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. *Annu Rev Plant Biol* 55: 459-493

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion analysis of variance. *J Am Stat Assoc* 47: 583-621

Liu K, Muse SV (2005) PowerMarker: integrated analysis environment for genetic marker data. *Bioinformatics* 21:2128-2129

Magalhaes JV, Garvin DF, Wang Y, Sorrells ME, Klein PE, Schaffert RE, Li L, Kochian LV (2004) Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the Poaceae. *Genetics* 167:1905-1914

Magalhaes JV, Liu J, Guimaraes CT, Lana UGP, Alves VMC, Wang YH, Schaffert RE, Hoekenga OA, Piñeros MA, Shaff JE, Klein PE, Carneiro NP, Coelho CM, Trick AN, Kochian LV (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nature Genetics* 39:1156-1161

Menkir A, Goldsbrough P, Ejeta G (1997) RAPD based assessment of genetic diversity in cultivated races of sorghum. *Crop Sci* 37:564-569

Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE (2004) Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. *Crop Sci* 44:1236-1244

Miklas PN, Delorme R, Hannan R, Dickson MH (1999) Using a subsample of the core collection to identify new sources of resistance to white mold in common bean. *Crop Sci* 39:569-573

Nei M, Li WH (1979) Mathematical models for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* 76:5269–5273

Parker DR, Norvell WA, Chaney RL (1995) GEOCHEM-PC: a chemical speciation program for IBM and compatible computers. In: Loeppert RH, Schwab AP, Goldberg S (eds) *Chemical equilibrium and reaction models*. Soil Science Society of America, Madison, pp 253–269

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959

Reddy BVS, Ramesh S, Reddy PS (2006) Sorghum genetic resources, cytogenetics and improvement. In: Singh RJ and Jauhar PP (ed) *Genetic resources, chromosome, engineering and crop improvement, vol cereals*, Taylor and Francis group, Boca Raton, pp 309-363

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4: 406-425

Saghai-Maroo MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer length polymorphism in barley: Mendelian inheritance, chromosomal location and population dynamics. *Proc Natl Acad Sci USA* 81: 8014-8018

Sanchez PA, Palm CA, Buol SW (2003) Fertility capability soil classification: a tool to help assess soil quality in the tropics. *Geoderma* 114:157-185

Scott AJ, Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 30:507-512

Snowden JD (1936) *The cultivated races of sorghum*. Adlard, UK

Tao Y, Manners JM, Ludlow MM, Henzell RG (1993) DNA polymorphisms in grain sorghum (*Sorghum bicolor* (L) Moench). Theor Appl Genet 86:679-688

Wang MX, Zhang HL, Zhang DL, Qi YW, Fan ZL, Li DY, Pan DJ, Cao YS, Qiu ZE, Yu P, Yang QW, Wang XK, Li ZC (2008) Genetic structure of *Oryza rufipogon* Griff. in China. Heredity 101:527-535

7. Tables and Figures

Table 1: Principal Component (PC) Analysis for $RNRG_{3d}$, $RNRG_{5d}$, VRD and b index (standardized variables). The dimension reduction achieved with the two first PCs, which explain 82.8 and 12.9% of the total variance, was deemed adequate to represent the total variability in the experimental space. Eigenvalues and the cumulative proportion of total variance (%) explained are shown.

Variables	PC1	PC2	PC3	PC4
$RNRG_{3d}$	0.48	-0.63	0.16	0.59
$RNRG_{5d}$	0.52	-0.34	-0.11	-0.77
VRD	0.49	0.54	0.68	-0.01
b index	0.50	0.44	-0.70	0.24
Eigenvalue	3.31	0.51	0.13	0.04
Cumulative Variance	82.8	95.7	98.9	100

Table 2: Aluminum tolerance in parents and derived F_{1s}: Percent Relative Root Growth (RRG) of F₁ hybrids and the respective parental lines grown in nutrient solution at {27} $\mu\text{M Al}^{3+}$. RRG values represent means of three replications (7 plants per replication). Genotypes whose RRG means are followed by the same lower-case letters are not statistically different by the t test (P<0.05). Marker - trait associations in backcross families: χ^2 test for goodness-of-fit to an expected 1:1 segregation ratio on backcrosses families and Single-marker analysis for M181, CTG29, S17 and S73 and RRG mean values. Average RRG for each marker class is shown. Within each backcross, genotypes classes whose RRG means are followed by the same lower-case letters are not statistically different by the t test (P<0.05). N: number of individuals for each marker class, Tt: heterozygous class and tt: homozygous sensitive class.

Al tolerance in parents and derived F _{1s}			Marker - trait associations in backcross families							
Genotypes	%RRG		Marker	N	χ^2	p-value (%)	F	p-value (%)	%RRG	
BR007B	14.38 $\pm$ 0.51	a	S17	37	9.15	0.25	10.72	0.14	Tt=17.30	a
IS10801	89.01 $\pm$ 13.62	b		68					tt=15.28	b
F ₁ (BR007B*IS10801)	18.75 $\pm$ 1.77	a	S73	39	9.31	0.23	11.08	0.12	Tt=17.28	a
				71					tt=15.28	b
BR007B	14.38 $\pm$ 0.51	a	S17	50	0.59	44.14	85.22	0	Tt=24.98	a
IS14351	114.31 $\pm$ 10.9	c		58					tt=12.45	b
F ₁ (BR007B*IS14351)	51.37 $\pm$ 3.77	b	S73	50	0.91	34.04	101.17	0	Tt=25.0	a
				60					tt=12.5	b
BR007B	14.38 $\pm$ 0.51	a	S17	17	3.60	5.79	0.43	100	Tt=18.89	a
IS15148	82.95 $\pm$ 7.59	c		30					tt=18.07	a
F ₁ (BR007B*IS15148)	22.93 $\pm$ 2.04	b								
BR007B	14.38 $\pm$ 0.51	a	S17	32	18.58	0.0016	2.1	15.06	Tt=21.44	a
IS15443	74.41 $\pm$ 3.15	c		77					tt=20.44	a
F ₁ (BR007B*IS15443)	21.11 $\pm$ 2.10	b								
BR007B	14.38 $\pm$ 0.51	a	S17	51	2.43	11.91	0.12	100	Tt=21.96	a
IS16396	76.46 $\pm$ 4.05	c		68					tt=22.28	a
F ₁ (BR007B*IS16396)	27.79 $\pm$ 2.12	b	S73	43	1.96	16.15	0.32	100	Tt=21.91	a
				57					tt=22.49	a
BR007B	14.38 $\pm$ 0.51	a	S17	26	1.33	24.92	5.72	2.0	Tt=20.08	a
IS21519	63.04 $\pm$ 9.79	b		35					tt=16.23	b
F ₁ (BR007B*IS21519)	23.47 $\pm$ 2.60	a	S73	28	0.41	52.21	10.03	0.24	Tt=21.04	a
				33					tt=15.88	b
BR007B	14.38 $\pm$ 0.51	a	CTG29	29	8.58	0.34	0.49	100	Tt=22.14	a
IS21622	31.65 $\pm$ 3.41	b		56					tt=22.86	a
F ₁ (BR007B*IS21622)	19.04 $\pm$ 1.45	a								
BR007B	14.38 $\pm$ 0.51	a	CTG29	51	0.26	61.17	0.45	100	Tt=16.08	a

IS23142	87.82 ± 6.38	b		46					tt=15.63	a
F ₁ (BR007B*IS23142)	23.20 ± 2.19	a								
BR007B	14.38 ± 0.51	a	S17	52	0.33	56.73	1.32	25.37	Tt=18.81	a
IS23644	28.96 ± 3.05	c		58					tt=17.88	a
F ₁ (BR007B*IS23644)	20.68 ± 1.28	b								
			S73	48	0.04	83.99	0.82	100	Tt=18.77	a
				50					tt=17.98	a
BR007B	14.38 ± 0.51	a	S17	26	083	36.21	4.01	4.99	Tt=18.73	a
IS23645	33.56 ± 2.20	c		33					tt=16.94	b
F ₁ (BR007B*IS23645)	24.71 ± 2.19	b								
			S73	25	0.08	78.15	3.60	6.34	Tt=18.88	a
				27					tt=17.00	a
BR007B	14.38 ± 0.51	a	S17	48	0.55	45.81	11.05	0.13	Tt=19.25	a
IS23666	46.03 ± 2.85	c		41					tt=15.78	b
F ₁ (BR007B*IS23666)	21.61 ± 2.01	b								
			S73	49	1.14	28.64	9.54	0.27	Tt=19.1	a
				39					tt=15.82	b
BR007B	14.38 ± 0.51	a	S17	61	11.38	0.74	15.00	0.2	Tt=13.70	a
IS24139	35.53 ± 4.41	b		29					tt=11.28	a
F ₁ (BR007B*IS24139)	21.11 ± 2.22	a								
BR007B	14.38 ± 0.51	a	S73	34	0.22	63.74	1.87	17.58	Tt=17.38	a
IS26457	56.74 ± 7.71	b		38					tt=15.29	a
F ₁ (BR007B*IS26457)	18.53 ± 2.52	a								
BR007B	14.38 ± 0.51	a	S17	40	3.64	5.62	25.29	0	Tt=18.23	a
IS26554	91.92 ± 11	c		59					tt=15.58	b
F ₁ (BR007B*IS26554)	25.98 ± 2.52	b								
			S73	40	4.75	2.94	27.02	0	Tt=18.23	a
				62					tt=15.53	b
BR007B	14.38 ± 0.51	a	S17	21	5.4	2.01	3.92	5.25	Tt=20.19	a
IS27490	22.16 ± 5.63	a b		39					tt=17.97	a
F ₁ (BR007B*IS27490)	16.47 ± 0.91	a b								
			S73	22	3.81	5.08	3.30	7.44	Tt=20.09	a
				37					tt=18.05	a
BR007B	14.38 ± 0.51	a	CTG29	43	0.27	60.02	0.04	100	Tt=17.02	a
IS29691	33.74 ± 3.63	c		48					tt=16.85	a
F ₁ (BR007B*IS26691)	24.73 ± 1.38	b								
BR007B	14.38 ± 0.51	a	S17	57	0.009	92.51	0.63	100	Tt=19.75	a
SSM215	19.17 ± 2.98	b		56					tt=18.68	a
F ₁ (BR007B*SSM215)	18.66 ± 1.73	a b								
			S73	56	0	100	0.66	100	Tt=19.78	a
				56					tt=18.68	a
BR007B	14.38 ± 0.51	a	CTG29	55	0.36	85.01	0.376	100	Tt=20.84	a
SSM1057	21.98 ± 1.37	b		57					tt=21.02	a
F ₁ (BR007B*SSM1057)	18.18 ± 1.11	a b								
			M181	56	0	100	0.13	100	Tt=20.80	a
				56					tt=21.14	a

Table 3: Analysis of six sorghum subpopulations for the Al tolerance variables Relative Net Root growth (*RNRG*) after 3 and 5 days of Al exposure, Visual Root Damage (*VRD*) and induction response (b index). Subpopulation mean scores of followed by the same lower-case letters are not statistically different by the non-parametric least significant difference (*lsd*) test ($P<0.05$). N: Number of individuals within each subpopulation.

Sub Populations	N	<i>RNRG</i> _{3d}			<i>RNRG</i> _{5d}			<i>VRD</i>		b index	
Q1	47	117.7	b	c d	179.3	a		176.8	a	172.3	a
Q2	58	103.9	b	c d e	106.6	b	c d	132.1	b c	122.0	b
Q3	23	144.3	a	b	134.3	a	b	153.8	a	116.6	b
Q4	43	119.4	b	c	124.2	b	c	99.8	b c d	131.4	a b
Q5	54	71.6		e	72.2		d	81.8		90.0	b
Q6	29	196.1	a		188.0	a		143.8	a b	137.6	a

Q1, guinea accessions from western Africa and guinea margaritifera; Q2, caudatum accessions from Africa and group of transplanted caudatum and durra accessions from Lake Chad region; Q3, Lines from Brazil and USA; Q4, kafir accessions from southern Africa; Q5, durra, bicolor and caudatum accessions from eastern Africa and Asia; Q6, guinea accessions from southern Africa and Asia.

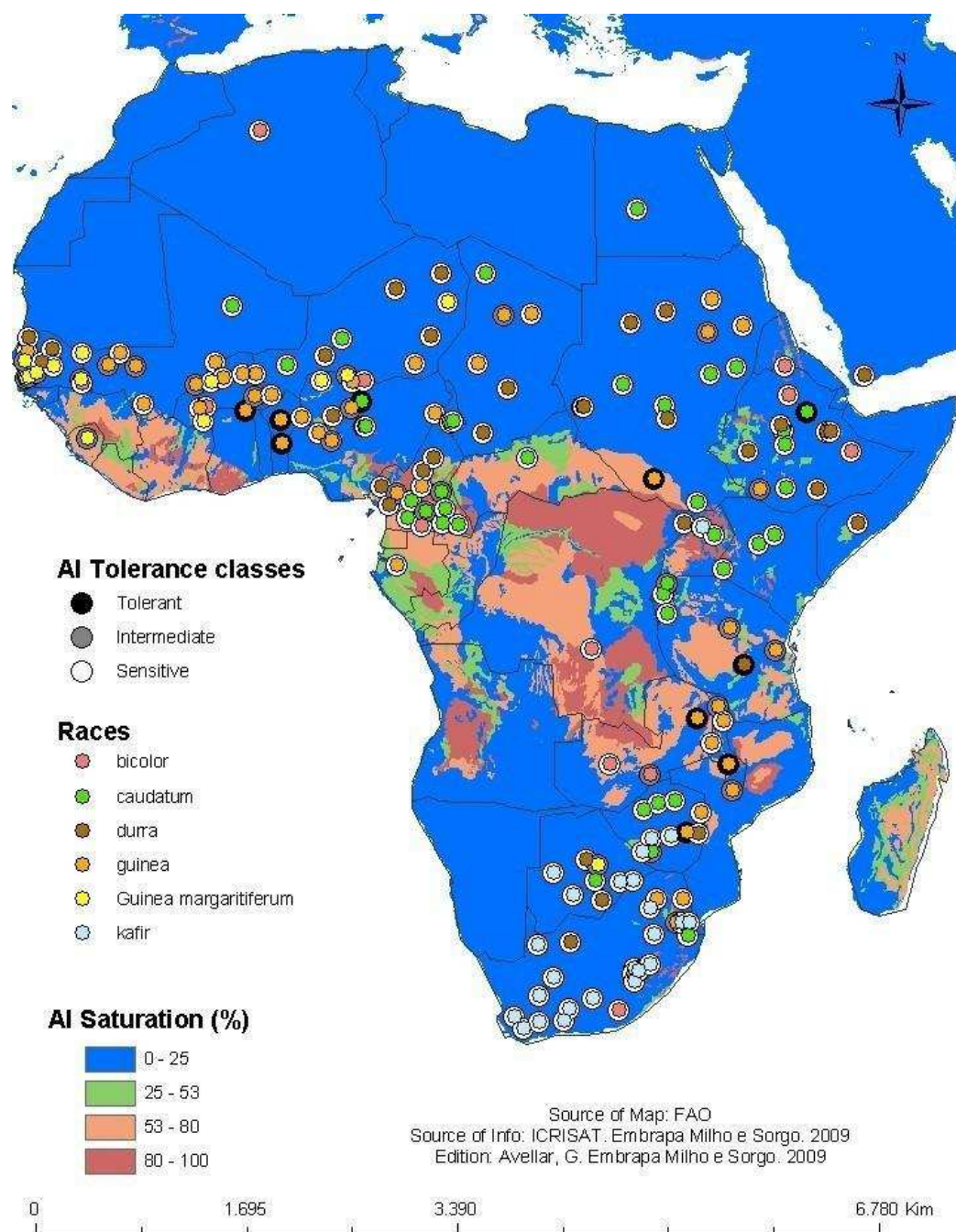


Figure 1: Geographical and racial distributions of the sorghum accessions evaluated in this study with respect to the occurrence of Al toxicity in Africa (Sanchez et al. 2003). Accessions were plotted on the map based on the latitude and longitude coordinates found in <http://www.icrisat.org/sorghum/Project1/pfirst.asp> (when this information was available). Accessions lacking those coordinates were plotted randomly within the known geographic origin (country). Outer circles indicate the respective classes of Al tolerance whereas inner circles indicate racial classification according to Harlan and de Wet (1972). Racial classifications can be found in (Deu et al. 2006) and www.ars-grin.gov.

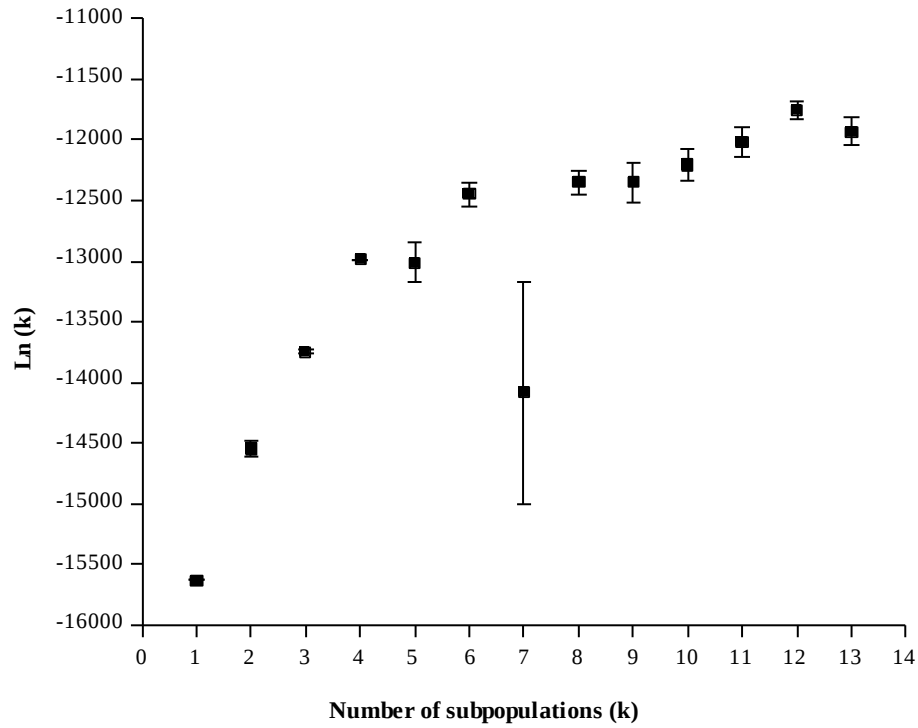


Figure 2: Ln of probability of data, $\text{Ln}(k)$, values for each number of subpopulation, k , obtained in STRUCTURE. Simulations were carried out with k varying from 1 to 13. $\text{Ln}(k)$ values are means of five independent runs for each k .

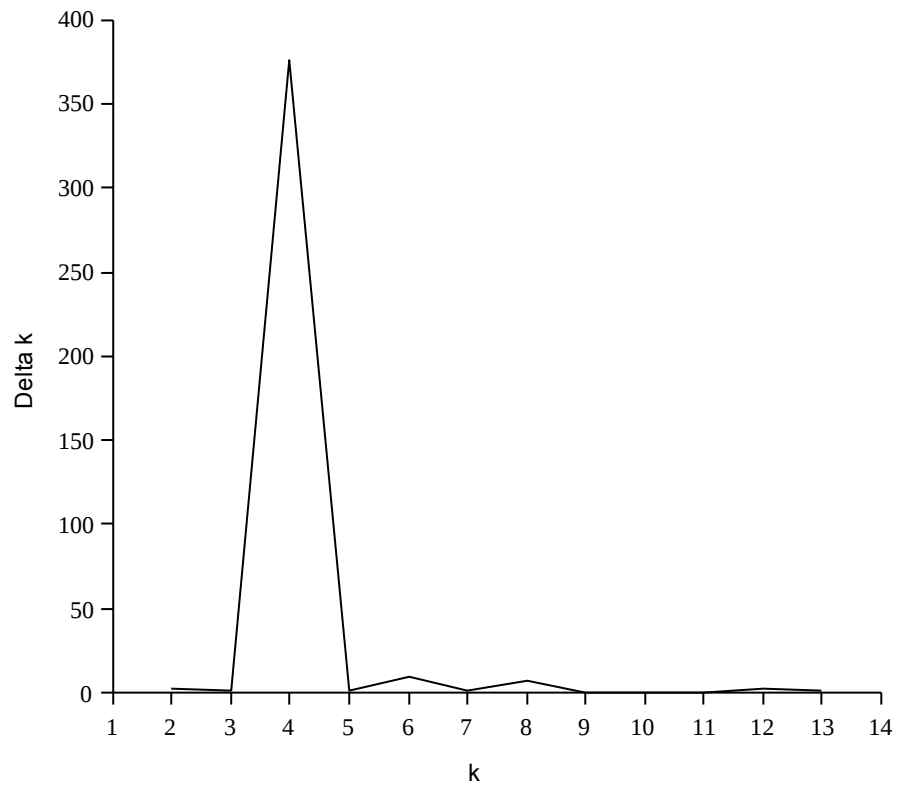


Figure 3: Plot of second order of change of Ln of probability of data, Delta k, (Evanno et al. 2005) values for different subpopulation numbers (k).

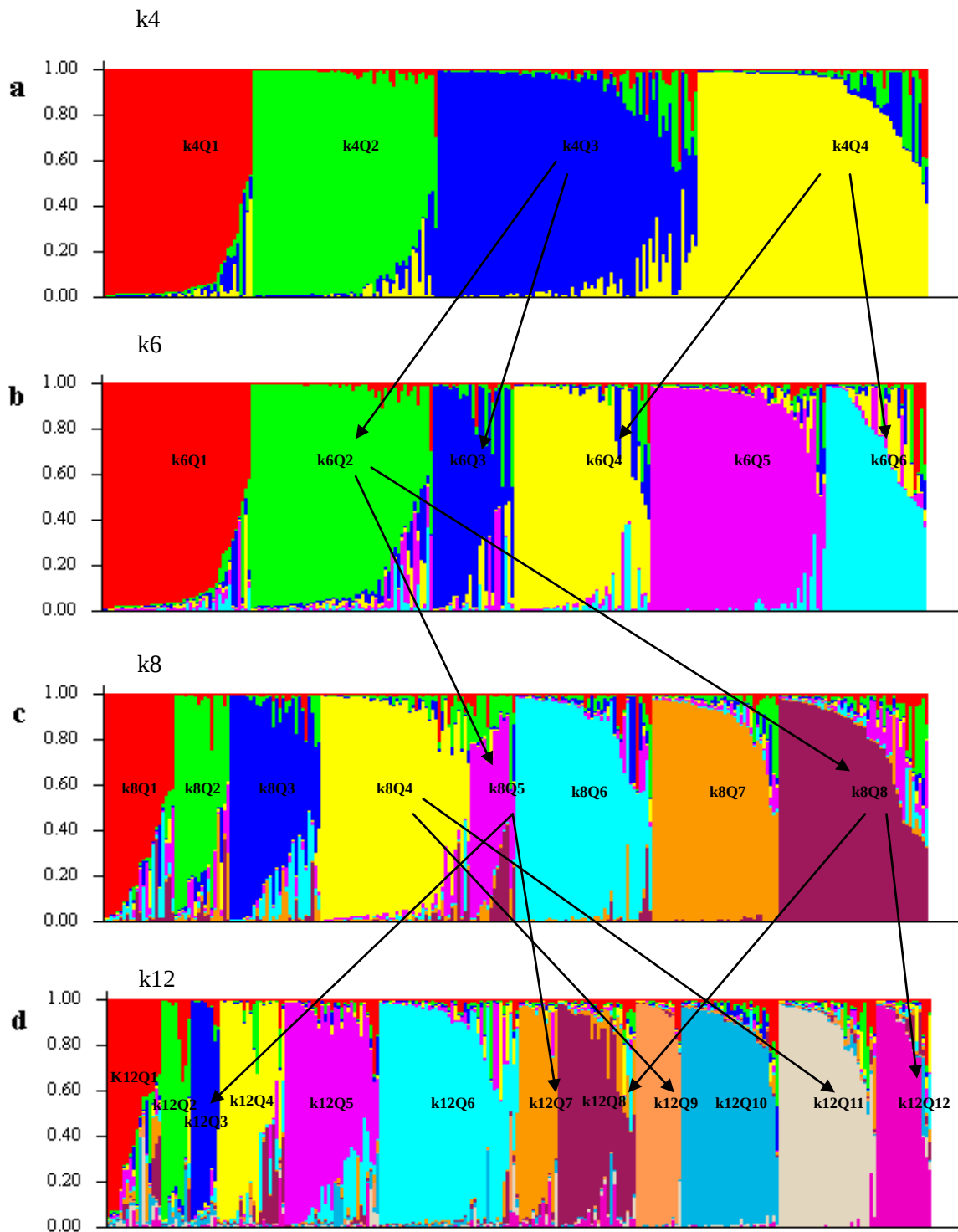


Figure 4: Membership of individual sorghum accessions to subpopulations based on 38 SSR marker loci for number of subpopulations, $k=4$ (a), $k=6$ (b), $k=8$ (c) and $k=12$ (d). Identity for the accessions attributed to subpopulations at the different k values are shown in Supplementary Table 5. Arrows indicate hierarchical subpopulation splits from $k=1$ to $k=12$.

Table S1: *Sorghum bicolor* accessions evaluated in this study. Country of origin, racial classification according to Harlan and de Wet (1972) as found in Deu et al (2006) followed by the variables $RNRG_{3d}$, $RNRG_{5d}$, VRD and b index. Values are means of two replications (7 plants per replication). Means followed by the same lower-case letters constitute homogeneous groups by the Scott-Knott test ($P<0.05$). CV: coefficient of variation.

Accession Name	Origin Country	ICRISAT	CIRAD	$RNRG_{3d}$		$RNRG_{5d}$		VRD		b index	
IS14351	MWI	G	G	233±7.6	j	262±4.6	m	5±0.2	d	14±1.2	h
SC283	TZA		G	105±5.4	i	157±9.0	l	4±0.3	d	9.3±1.3	f
CMSXS226R	BRA			118±5.7	i	163±6.1	l	4±0	d	11±0.9	g
CMSXS227	BRA			116±5.9	i	139±4.8	j	5±0.2	d	12±0.8	h
IS21519	MWI	G	G	95±5.2	h	113±7.9	i	4±0.8	d	8.7±0.1	f
IS23142	TZA	D	D	86±5.9	h	163±9.9	i	5±0.3	d	11.1±2.3	g
IS25077	GHA	G	G	118±5.9	i	120±8.9	i	3±0.4	c	7.2±1.2	e
9929034				82±3.6	h	113±5.4	i	5±0.3	d	16±1.8	i
IS26457	BEN	G	G	96±6.1	h	97±9.3	h	4±0.2	d	12.8±1.8	h
IS26554	BEN	G	G	74±3.1	h	95±4.3	h	5±0.2	d	20.3±2.8	j
IS29691	ZWE	G	G	117±5.6	i	99±6.5	h	1±0.3	a	2.7±0.5	b
SC566-14	NGA		C	97±5.8	h	99±3.9	h	5±0	d	13.3±0.7	h
IS10801	TCD	GC	GC	65±4.9	g	78±6.5	g	4±0.5	d	9.6±0.5	f
5DX61-6-2	UGA			76±4.1	h	89±6.3	g	4±0.3	d	11±2.6	g
BRANDES	USA			69±4.7	g	82±8.5	g	5±0.2	d	9.8±0.6	f
SC175-14	ETH		C	72±2.4	h	87±3.0	g	4±0.3	c	19.2±0.7	j
CMSXS225R				88±7.0	h	85±7.9	g	3±0.3	c	7.8±2.1	e
ATF14	BRA			79±2.9	h	77±6.6	g	2±0.3	b	4.8±0.2	d
IS24139	TZA	G	G	60±2.1	f	66±6.6	f	4±0.3	d	9.1±1.2	f
IS27490	BFA	G	G	89±5.2	h	67±2.2	f	2±0.2	b	2.1±0.2	b
9DX9-11	UGA			57±2.5	e	61±4.2	f	4±0.3	d	8±0.7	e
SC549	NGA		G	70±3.4	g	61±4.9	f	3±0.3	b	6±0.9	e
IS15148	CMR	C	C	44±1.0	d	51±2.0	e	4±0.4	d	11±0.8	g
IS15443	CMR	C	C	44±2.6	d	52±3.4	e	4±0.4	d	11±0	g
IS21622	MWI	G	G	107±9.3	i	53±4.5	e	2±0.3	b	0.4±0	a
IS21849	SLE	G	Gm	40±0.7	c	51±1.2	e	4±0.6	d	7.7±0.5	e
IS23666	GMB	G	Gm	53±3.7	d	53±4.1	e	3±0.3	c	7.3±0.5	e
IS23777	MWI	G	G	76±1.9	h	57±2.3	e	2±0.3	b	2.8±0.4	c
IS23100	TZA	G	G	76±6.6	g	55±8.4	e	1±0.3	a	2.7±0.5	b
IS24072	TZA	G	G	60±2.9	f	55±2.9	e	4±0.7	d	7.5±0.5	e
3DX-57-1-1-9-D	UGA			53±3.0	d	55±4.9	e	3±0.7	c	6.5±2.5	e
IS3958	NPL	G	G	79±2.8	h	44±2.2	d	1±0.5	a	0.6±0.4	a

IS12931	CHN	G	G	76±5.2	h	40±2.4	d	1±0.3	a	0.4±0.1	a
IS16396	CMR	G	G	34±1.1	c	44±2.4	d	4±0.7	c	9.4±1.8	f
IS19455	BWA	G	Gm	66±2.5	g	40±1.5	d	2±0.3	b	0.5±0.1	a
IS20016	SEN	G	G	49±1.5	d	44±1.0	d	3±0.7	b	5±0.2	d
IS20064	SEN	G	Gm	51±2.2	e	40±1.8	d	3±0.3	b	1.9±0	b
IS23644	GMB	G	Gm	42±1.2	c	44±1.7	d	3±0.3	c	9±0.5	f
IS23645	GMB	G	Gm	42±2.8	c	44±3.6	d	4±0.7	c	7±0.5	e
IS23669	GMB	G	Gm	45±1.9	d	46±2.9	d	3±0.3	c	7.6±0.1	e
IS25596	RWA	C	C	43±2.3	d	46±4.9	d	4±0.6	d	10±2.9	f
IS19466	LKA	G	G	42±2.4	c	33±2.2	c	1±0.2	a	0.5±0.1	a
IS19685	IND	G	G	52±3.6	e	33±1.9	c	1±0.2	a	0.4±0	a
IS22282	BWA	B	B	51±3.7	e	35±2.8	c	1±0.2	a	1.7±0.3	b
IS26041	MLI	G	G	60±5.2	f	33±2.2	c	1±0	a	0.5±0.5	a
IS27390	BFA	G	G	51±3.2	e	33±4.2	c	2±0.4	b	2.9±1.1	c
IS29876	ZWE	GC	C	46±4.2	d	38±4.7	c	2±0.5	b	4±1.6	c
CMSXS235	BRA			40±2.7	c	34±3.3	c	1±0.3	a	2.7±0.9	b
IS2430	NGA	G	Gm	34±1.3	c	27±0.8	b	2±0.2	b	2±0	b
IS2848	ZAF	C	C	31±2.4	b	21±1.7	b	1±0.2	a	0.8±0.2	a
IS3780	CHN	GB	B	36±2.6	c	24±1.7	b	2±0.3	b	0.8±0	a
IS3905	MLI	G	G	45±2.8	d	27±2.2	b	2±0.2	b	0.4±0.4	a
IS3957	NPL	G	G	45±4.2	d	27±3.0	b	1±0	a	0.9±0.5	a
IS3959	NPL	G	G	53±2.8	e	31±2.0	b	2±0.3	b	0.7±0.1	a
IS3967	IND	G	G	45±2.2	d	30±1.6	b	2±0	b	1±0.4	a
IS4963	IND	G	G	31±1.9	b	21±1.5	b	1±0	a	0.3±0.1	a
IS6745	BFA	G	G	38±1.1	c	21±0.7	b	1±0	a	0.3±0	a
IS7861	NGA	G	G	34±1.4	c	31±1.8	b	3±0.2	c	6.4±0.2	e
IS7889	NGA	G	G	39±1.9	c	26±1.9	b	2±0.6	b	0.8±0.6	a
IS9597	NER	G	Gm	35±2.1	c	30±2.0	b	1±0.2	a	21±0.2	b
IS10194	BFA	GB	B	39±1.4	c	26±0.8	b	2±0.2	b	0.7±0.1	a
IS10876	NGA	GC	C	27±1.2	b	23±1.5	b	2±0	b	3.3±0.6	c
IS11119	ETH	C	C	30±1.1	b	20±0.7	b	1±0	a	1±0.2	a
IS12179	ETH	B	B	31±2.7	b	22±1.4	b	1±0.2	a	1.3±0.1	a
IS12531	ETH	B	B	27±2.1	b	20±1.3	b	1±0.2	a	0.8±0.1	a
IS13989	ZAF	K	K	23±1.1	b	22±1.4	b	1±0.2	a	2.3±0.6	b
IS14317	SWZ	G	G	43±2.6	d	26±1.5	b	1±0	a	0.1±0	a
IS14331	ZAF	G	G	61±3.3	f	30±1.7	b	1±0.2	a	0.4±0.1	a
IS14417	MWI	G	G	38±1.3	c	21±0.6	b	1±0.2	a	0.3±0.2	a
IS16044	CMR	C	C	29±1.1	b	23±1.5	b	2±0.2	b	3±0.7	c
IS17658	GHA	G	G	31±1.1	b	26±1.4	b	3±0.6	c	3±0.3	c
IS20097	SEN	G	G	30±0.7	b	23±0.7	b	2±0.4	b	2.3±0.1	b
IS22287	BWA	K	K	29±1.6	b	20±1.2	b	1±0	a	1.5±0.2	b

IS22330	BWA	D	D	27±0.9	b	21±0.9	b	1±0	a	1.6±0.1	b
IS22332	BWA		KC	30±1.2	b	21±0.9	b	1±0	a	1.2±0	a
IS22334	BWA	K	K	33±1.7	b	27±1.7	b	1±0	a	2.4±0.4	b
IS22893	SDN	GC	GC	39±1.4	c	31±1.2	b	1±0	a	2.9±0	c
IS23178	ZMB	B	B	35±2.7	c	31±2.4	b	1±0.2	a	0.9±0.3	a
IS23254	ZMB	B	B	37±2.7	c	28±1.8	b	1±0.3	a	0.5±0	a
IS24481	ZAF	K	K	56±4.8	e	29±1.9	b	2±0.3	b	0.7±0	a
IS24503	SEN			32±1.1	b	22±0.8	b	1±0.3	a	0.7±0.1	a
IS24887	NGA	G	G	33±1.0	b	20±0.5	b	2±0.2	b	1.1±0.1	a
IS25499	BDI	C	C	31±1.2	b	23±1.1	b	2±0.2	b	2.6±0.3	b
IS25702	MLI	GC	GC	35±1.8	c	26±1.7	b	2±0.2	b	2.4±0	b
IS25733	MLI	G	G	34±2.8	c	25±2.3	b	2±0.3	b	1.8±0.2	b
IS26110	MLI	G	G	49±2.2	d	29±1.3	b	2±0.2	b	0.8±0.1	a
IS27146	ZWE	D	D	32±1.4	b	21±0.9	b	1±0.2	a	0.8±0.2	a
IS28645	YEM	DC	DC	34±2.0	c	24±1.3	b	1±0.3	a	1.1±0.1	a
IS29233	SWZ	K	K	30±1.1	b	27±1.4	b	2±0.4	b	1.8±0.4	b
IS29310	SWZ	GC	KC	29±1.6	b	20±1.3	b	1±0.3	a	0.9±0.1	a
IS29375	LSO	GC	GC	30±1.5	b	21±1.0	b	1±0.2	a	1±0	a
IS29407	LSO	K	KC	39±1.5	c	24±0.9	b	1±0	a	0.9±0	a
IS29409	LSO	K	K	37±2.1	c	27±1.5	b	1±0.2	a	0.8±0	a
IS29606	ZAF	K	K	31±1.8	b	22±1.5	b	1±0.2	a	1.5±0.1	b
IS29872	ZWE	C	KC	29±1.1	b	24±1.1	b	1±0.3	a	2±0.3	b
IS29911	ZWE	C	KC	27±0.7	b	20±0.6	b	1±0.3	a	2.2±0.2	b
IS30443	CHN	C	CB	30±2.3	b	21±1.6	b	1±0.2	a	1±0	a
IS31559	BDI	C	C	31±2.5	b	25±1.5	b	1±0.2	a	2.4±0.6	b
SSM232	BFA		G	44±3.8	d	25±2.1	b	1±0.3	a	0.6±0.1	a
SSM249	BFA		G	28±1.3	b	20±1.1	b	2±0.2	b	1.9±0.1	b
SSM261	BFA		Gm	48±2.0	d	27±0.9	b	1±0.3	a	0.6±0.1	a
SSM275	BFA		Gm	32±1.3	b	23±1.1	b	3±0.4	c	1.4±0.1	a
SSM276	BFA		G	29±1.9	b	23±1.4	b	2±0	b	1.5±0.3	b
SSM501	NER		Gm	25±0.5	b	24±1.1	b	2±0.3	b	2.6±0.6	b
SSM505	NER		D	35±1.1	c	28±1.2	b	3±0.4	c	2.6±0.7	b
SSM1103	TCD		C	28±1.5	b	24±1.4	b	2±0.3	b	3.1±0.3	c
SC112-14	ETH		C	38±4.3	c	24±2.9	b	1±0	a	0.3±0.1	a
BR001B	USA			31±1.0	b	23±0.9	b	1±0.2	a	2.1±0.4	b
IS8577				35±2.3	c	26±2.1	b	1±0.2	a	1.6±0.2	b
BTX631	USA		KC	29±3.3	b	21±2.3	b	2±0.2	b	1±0.1	a
V20-1-1	UGA			43±3.8	d	29±3.5	b	1±0.2	a	1.5±0.3	b
9929020	BRA			33±3.9	b	29±3.7	b	2±0.2	b	3.2±0.1	c
ARG1	USA/ARG			48±4.7	d	27±2.7	b	2±0	b	0.3±0	a
9929030	BRA			28±2.1	b	25±2.3	b	2±0.5	b	3.8±1.4	c

CMSXS182	BRA			36±1.3	c	25±1.4	b	1±0.3	a	1±0.2	a
IS13	USA	B	B	22±0.8	a	15±0.6	a	2±0.2	b	0.4±0	a
IS303	CHN	KB	B	28±1.9	b	18±1.1	a	1±0.3	a	0.4±0	a
IS929	SDN	D	D	12±0.7	a	9±0.4	a	1±0.2	a	0.8±0.1	a
IS1398	SDN	D	D	21±1.1	a	15±0.9	a	1±0.3	a	0.2±0.1	a
IS2156	NGA	B	B	23±0.6	b	15±0.6	a	2±0.3	b	0.5±0	a
IS2262	SDN	KC	C	13±0.7	a	8±0.4	a	1±0.2	a	0.1±0.1	a
IS2263	SDN	D	D	13±0.4	a	9±0.3	a	1±0	a	0.4±0.2	a
IS2398	ZAF	K	K	27±1.5	b	17±1.1	a	2±0.2	b	0.5±0	a
IS2416	ZAF	K	K	25±0.6	b	16±0.4	a	1±0.2	a	0.6±0.2	a
IS2787	KEN	C	C	27±0.9	b	16±0.6	a	2±0.2	b	0.1±0	a
IS2807	ZWE	C	C	22±0.5	a	15±0.5	a	2±0.2	b	0.5±0	a
IS2814	ZWE	C	C	23±1.0	b	14±0.7	a	1±0.2	a	0.1±0	a
IS3073	SDN	C	C	22±1.0	a	14±0.7	a	1±0	a	0.4±0.1	a
IS3421	IND	G	G	26±2.0	b	17±1.4	a	2±0.2	b	0.5±0.3	a
IS3511	SDN	KC	C	21±1.2	a	13±0.7	a	1±0	a	0.2±0	a
IS3771	CHN	C	CB	20±0.8	a	13±0.5	a	1±0.2	a	0.2±0	a
IS3971	IND	D	D	19±0.8	a	13±0.7	a	1±0.2	a	0.3±0.2	a
IS4027	IND	B	B	18±1.1	a	13±0.9	a	1.0±0	a	0.9±0.3	a
IS4285	IND	D	D	21±1.4	a	16±1.2	a	1±0.2	a	1.3±0.3	a
IS4821	IND	D	D	18±1.0	a	12±0.6	a	2±0.2	b	0.5±0.2	a
IS5430	IND	B	B	23±1.1	b	16±0.7	a	1±0	a	0.2±0.1	a
IS5867	IND	DC	DC	13±0.5	a	9±0.3	a	1±0	a	0.4±0	a
IS5972	IND	D	D	20±0.9	a	14±0.7	a	1±0	a	1.2±0.4	a
IS6118	IND	DC	DC	29±2.2	b	19±1.4	a	1±0.3	a	0.7±0.1	a
IS6193	IND	D	D	14±0.6	a	9±0.4	a	1±0	a	0.2±0.2	a
IS6351	IND	D	D	15±0.8	a	11±0.4	a	1±0.2	a	0.3±0.1	a
IS6828	BFA	G	G	29±1.4	b	19±0.9	a	1±0	a	0.4±0.1	a
IS7287	NGA	C	G	6±0.6	a	5±0.4	a	1±0	a	0.2±0.1	a
IS8685	SDN	C	C	14±0.7	a	10±0.5	a	1±0	a	0.3±0.1	a
IS8882	UGA	C	C	22±0.8	a	17±0.6	a	1±0.3	a	1±0	a
IS9303	ZAF	K	K	25±2.0	b	18±1.6	a	1±0	a	0.6±0	a
IS9331	ZAF	K	C	26±1.3	b	19±0.9	a	1±0	a	0.3±0	a
IS9468	ZAF	K	K	21±0.8	a	15±0.5	a	1±0	a	0.7±0.1	a
IS9527	ZAF	K	K	23±1.3	b	17±1.2	a	1±0.2	a	1.5±0	b
IS10234	CAF	GC	C	18±1.3	a	13±1.3	a	2±0.2	b	1.6±1.2	b
IS10844	TCD	G	GC	29±2.5	b	18±1.5	a	2±0.2	b	0.9±0.2	a
IS10882	NGA	GC	C	23±1.3	b	15±0.8	a	2±0.2	b	0.4±0.3	a
IS11026	ETH	D	D	23±1.1	b	17±1.1	a	2±0	b	2.0±0.6	b
IS11827	ETH	D	D	17±1.1	a	13±0.6	a	1±0	a	0.7±0.2	a
IS12169	ETH	B	B	25±1.0	b	16±0.5	a	1±0.2	a	0.3±0.2	a

IS12447	SDN	DC	GC	23±1.5	a	15±0.8	a	2±0.2	b	0.7±0.3	a
IS12542	ETH	D	D	22±1.7	a	15±1.2	a	1±0	a	0.8±0.2	a
IS12804	TUR	B	B	22±1.0	a	16±0.7	a	1±0.3	a	0.2±0.2	a
IS13113	IND	B	B	29±1.7	b	19±1.0	a	1±0	a	0.4±0.3	a
IS13452	ZWE	GC	C	24±1.0	b	16±0.6	a	1±0.3	a	0.2±0.1	a
IS13791	ZAF	K	KC	27±1.1	b	19±0.7	a	2±0.4	b	0.2±0	a
IS13845	ZAF	K	K	14±1.0	a	10±0.8	a	1±0.2	a	0.1±0	a
IS13848	ZAF	C	C	30±1.1	b	19±0.7	a	1±0.2	a	0.2±0.1	a
IS13926	ZAF	K	KC	24±1.4	b	20±1.1	a	1±0	a	0.5±0	a
IS14414	MWI	G	G	18±1.2	a	12±0.5	a	2±0.4	b	0.3±0.1	a
IS15752	CMR	C	C	25±2.2	b	16±1.3	a	2±0.2	b	0.6±0.1	a
IS16101	CMR	D	D	19±1.5	a	12±0.9	a	1±0.3	a	0.2±0.1	a
IS16125	CMR	C	C	21±0.7	a	15±0.8	a	2±0.3	b	0.1±0.1	a
IS16173	CMR	C	C	21±1.2	a	12±0.6	a	1±0.2	a	0.2±0	a
IS16186	CMR	C	D	16±0.7	a	11±0.6	a	1±0.3	a	0.9±0.3	a
IS16545	CMR	C	C	21±1.2	a	13±0.8	a	1±0.3	a	0.4±0.1	a
IS19026	SDN	C	DC	10±0.6	a	8±0.4	a	2±0.2	b	0.4±0.1	a
IS19132	SDN	GC	CB	18±1.1	a	12±0.8	a	2±0.5	b	0.3±0.1	a
IS19453	BWA	D	D	17±0.5	a	12±0.4	a	2±0.6	b	1.6±0	b
IS19847	IND	DC	DC	26±3.1	b	17±2.0	a	1±0.2	a	0.5±0.2	a
IS19953	SEN	G	Gm	25±0.9	b	17±0.7	a	1±0.3	a	0.5±0.2	a
IS20351	NER	D	D	15±1.2	a	10±0.7	a	1±0	a	0.1±0.1	a
IS20689	USA	B	B	23±1.8	b	15±1.1	a	1±0.3	a	0.1±0.1	a
IS20706	USA	B	B	28±1.1	b	19±0.8	a	2±0.4	b	0.1±0.1	a
IS20727	USA	B	B	21±0.7	a	16±0.6	a	1±0.3	a	0.4±0	a
IS20864	CHN	C	C	25±1.5	b	16±1.1	a	2±0.8	b	0.5±0.1	a
IS21124	KEN	C	C	24±0.5	b	15±0.4	a	2±0.6	b	0.4±0.1	a
IS21891	USA	C	C	27±1.8	b	18±1.1	a	2±0.2	b	0.3±0.1	a
IS22239	BWA	K	K	16±0.7	a	11±0.6	a	2±0.5	b	0.1±0.1	a
IS22294	BWA	K	K	26±1.0	b	19±0.7	a	2±0.2	b	1.3±0.1	a
IS24009	USA	C	C	20±0.7	a	14±0.6	a	2±0.4	b	1.2±0.1	a
IS26731	ZAF	B	B	29±1.4	b	19±0.9	a	1±0	a	0.6±0.1	a
IS26833	SDN	C	C	21±1.0	a	15±0.7	a	1±0.2	a	0.6±0.2	a
IS27164	ZWE	K	K	27±1.2	b	18±0.9	a	1±0	a	1.2±0.3	a
IS27891	ZAF	GC	GC	28±1.3	b	19±0.9	a	2±0.3	b	0.6±0.4	a
IS28409	YEM	D	D	19±1.5	a	13±1.2	a	2±0.6	b	1.5±0.3	b
IS29226	SWZ	GC	KB	20±1.1	a	12±0.5	a	1±0.3	a	0.1±0.1	a
IS29496	LSO	K	K	23±0.6	b	17±0.4	a	1±0	a	0.9±0.2	a
IS29569	LSO	K	K	31±0.9	b	18±0.5	a	1±0	a	0.3±0	a
IS30030	ZWE	K	K	25±0.7	b	18±0.7	a	1±0.2	a	0.6±0.2	a
IS30175	ZWE	G	G	28±2.1	b	18±1.2	a	1±0.2	a	0.4±0.1	a

IS30352	CHN	DC	CB	21±2.0	a	17±1.7	a	1±0	a	0.6±0.2	a
IS30385	CHN	C	C	18±0.6	a	12±0.4	a	1±0.2	a	0.5±0.1	a
IS30400	CHN	C	KC	21±1.2	a	13±0.7	a	1±0.2	a	0.5±0.1	a
IS30405	CHN	DC	BC	26±1.5	b	17±1.0	a	1±0	a	0.8±0.1	a
IS30417	CHN	DC	BC	25±4.7	b	14±2.1	a	1±0.2	a	0.2±0.2	a
IS30436	CHN	C	C	20±1.2	a	13±0.7	a	1±0	a	0.2±0.1	a
IS30441	CHN	C	C	25±1.6	b	15±0.9	a	1±0.2	a	0.3±0	a
IS30451	CHN	C	C	28±1.3	b	18±0.7	a	1±0.2	a	0.8±0	a
IS30538	KOR	B	B	16±0.9	a	10±0.4	a	1±0.2	a	0.2±0	a
IS31681	DZA	B	B	23±1.0	b	16±0.7	a	1±0.2	a	1.3±0.4	a
IS32569	SOM	D	D	16±1.0	a	10±0.6	a	1±0	a	0.3±0.1	a
IS33116	CMR	C	C	20±1.4	a	14±0.9	a	2±0.4	b	0.5±0.3	a
IS33261	CMR	C	C	24±1.1	b	15±0.5	a	1±0.2	a	0.5±0.1	a
IS33353	KEN	C	C	19±1.4	a	14±0.7	a	1±0	a	0.5±0.1	a
SSM12	CMR		D	20±1.7	a	14±1.2	a	1±0	a	0.3±0	a
SSM19	CMR		D	23±1.4	b	16±0.9	a	1±0.2	a	0.9±0.1	a
SSM29	CMR		D	14±1.0	a	9±0.7	a	1±0	a	0.4±0.1	a
SSM205	BFA		G	27±1.8	b	18±1.1	a	1±0	a	0.4±0.1	a
SSM215	ETH		C	26±0.8	b	19±0.7	a	4±0.3	c	2.5±0.2	b
SSM379	MLI		DC	23±1.7	b	15±1.0	a	1±0	a	0.4±0	a
SSM546	NER		C	13±0.6	a	8±0.4	a	1±0.3	a	0.1±0	a
SSM547	NER		C	21±1.3	a	15±0.6	a	1±0.3	a	1±0.2	a
SSM552	NER		G	28±0.9	b	19±0.6	a	2±0.3	b	0.4±0.1	a
SSM557	NER		DC	24±1.0	b	15±0.6	a	2±0.4	b	0.1±0.1	a
SSM625	NGA		D	32±1.7	b	19±1.0	a	2±0.3	b	0.8±0.1	a
SSM964	SEN		D	20±1.9	a	12±1.0	a	1±0.3	a	0.4±0.1	a
SSM973	SEN		G	11±0.7	a	7±0.4	a	1±0	a	0.4±0.2	a
SSM1046	SEN		B	24±1.4	b	17±1.1	a	2±0.3	b	1.7±0.1	b
SSM1049	SEN		Gm	25±0.9	b	16±0.5	a	2±0.4	b	0.3±0.2	a
SSM1057	SEN		DC	25±1.1	b	19±1.2	a	2±0.5	b	2±0.1	b
SSM1102	TCD		DC	20±1.6	a	13±1.2	a	1±0.5	a	0.6±0.3	a
SSM1123	NER		DC	23±1.3	b	15±0.9	a	1±0.5	a	0.3±0.1	a
SSM1267	CMR		B	23±1.1	b	17±1.0	a	2±0.5	b	1.7±0.4	b
SSM1284	ZAR		B	17±0.8	a	12±0.4	a	1±0	a	0.3±0.2	a
SSM1370	ZAF		D	20±1.0	a	12±0.5	a	1±0.2	a	0.1±0.1	a
SSM1592	TCD		D	17±0.9	a	12±0.7	a	2±0.2	b	0.5±0	a
SSM1611	TCD		D	21±1.7	a	14±1.1	a	1±0.2	a	0.5±0.1	a
SC114-12	SDN		KC	30±2.4	b	18±1.3	a	1±0.2	a	0.4±0.1	a
BTX623	USA			28±2.0	b	16±1.1	a	1±0.2	a	0.1±0.1	a
IS3620C	NGA		G	28±1.5	b	18±1.2	a	2±0.2	b	0.3±0	a
9910032	BRA			37±4.3	c	18±2.2	a	1±0.3	a	0.2±0.2	a

IS10622B			22±1.5	a	15±1.1	a	1±0.2	a	0.3±0.1	a
TX2785N	USA		20±0.5	a	13±0.3	a	1±0	a	0.2±0	a
TX2784	USA		21±1.1	a	14±0.9	a	1±0	a	0.1±0.1	a
CMSXS153	UGA		25±1.0	b	15±0.6	a	1±0.2	a	0.3±0.2	a
BR007B	USA		19±1.2	a	12±0.8	a	1±0	a	0.4±0.1	a
BR012R	BRA		24±1.8	b	16±1.2	a	1±0.3	a	0.3±0.1	a
BTX644	USA		18±1.4	a	11±0.9	a	1±0.2	a	0.1±0	a
TX430	USA		29±2.3	b	19±1.6	a	1±0	a	0.8±0.1	a
QL3	AUS		29±2.3	b	18±1.2	a	1±0.3	a	0.6±0.2	a
BTX643	USA		24±1.1	b	14±0.5	a	2±0	b	0.4±0.1	a
CMSXS116B	ETH	C	23±1.0	b	14±0.7	a	2±0.3	b	0.2±0.2	a
BTX642	ETH	D	26±1.9	b	15±0.9	a	1±0.2	a	0.1±0.1	a
IS10317			22±1.7	a	16±0.7	a	2±0.2	b	0.8±0.4	a
SC748	SDN	GC	20±0.5	a	13±0.3	a	1±0.2	a	0.1±0.1	a
CMSXS156	USA		30±1.2	b	19±0.9	a	2±0.2	b	0.4±0.1	a
CMSXS912	USA		25±2.7	b	14±1.5	a	1±0.2	a	0.2±0	a
CMSXS180	BRA		28±1.5	b	17±1.0	a	1±0.2	a	0.4±0.1	a
CMSXS230			26±1.3	b	16±0.8	a	1±0.3	a	0.5±0	a
CV(%)			11		19		27		39	

Racial classification performed in accordance with Harlan and de Wet (1972); B: bicolor; C: caudatum; D: durra; G: guinea; Gm: guinea margaritifera; K: kafir; DC: intermediate durra-caudatum; BC: intermediate bicolor-caudatum; GC: intermediate guinea-caudatum; KC: intermediate kafir-caudatum. Country codes are as follows: ARG, Argentina; BEN, Benin; BDI, Burundi; BFA, Burkina Faso; BRA, Brazil; BWA, Botswana; CAF, Central African Republic; CHN, China; CMR, Cameroon; DZA, Algeria; ETH, Ethiopia; GHA, Ghana; GMB, Gambia; IND, India; KEN, Kenya; KOR, South Korea; LKA, Sri Lanka; LSO, Lesotho; MLI, Mali; MWI, Malawi; NER, Niger; NGA, Nigeria; NPL, Nepal; RWA, Rwanda; SDN, Sudan; SEN, Senegal; SLE, Sierra Leone; SOM, Somalia; SWZ, Swaziland; TCD, Chad; TUR, Turkey; TZA, Tanzania; UGA, Uganda; USA, United States of America; YEM, Yemen; ZAF, Republic of South Africa; ZAR, Democratic Republic of the Congo (Zaire), ZMB; Zambia; ZWE, Zimbabwe.

Table S2: Relative percent values of net root growth inhibition (*RNRG*) of the sorghum accessions grown in nutrient solution at {39} $\mu\text{M Al}^{3+}$ and {60} $\mu\text{M Al}^{3+}$ after three and five days of Al stress, followed by the induction (b) index. Values are means of two replications (7 plants per replication). Lines whose means are followed by the same lower-case letters within each of the two Al activities constitute homogeneous groups by the Scott-Knott test ($P < 0.05$). CV: coefficient of variation.

Accessions	{39} $\mu\text{M Al}^{3+}$						{60} $\mu\text{M Al}^{3+}$					
	<i>RNRG</i> _{3d}		<i>RNRG</i> _{5d}		b index		<i>RNRG</i> _{3d}		<i>RNRG</i> _{5d}		b index	
IS14351	84±2.6	g	103±6.0	d	11±1.9	c	47±4.9	d	47±5.7	c	3±1.0	c
SC566-14	61±2.8	e	78±2.7	c	12±1.3	c	28±2.4	c	18±1.7	b	0.3±0.2	a
IS10801	55±1.3	e	76±2.2	c	9±1.0	c	25±1.6	c	25±2.6	b	2±1.0	b
IS23100	70±22.3	f	75±20.3	c	6±6	b	20±3.0	c	16±2.2	b	0.50±0	a
IS15443	45±4.0	d	46±6.1	b	7±0.3	b	20±2.2	c	16±2.5	b	1±0.5	a
IS21519	51±2.7	d	56±2.6	b	9±0.4	c	21±1.2	c	14±1.3	a	0.6±0.1	a
IS21849	48±1.8	d	61±2.2	b	9±0.3	c	21±1.5	c	21±1.6	b	2±0	b
IS23142	36±2.2	c	52±2.9	b	4±0	a	20±1.0	c	18±0.7	b	0.3±0	a
BR007B	13±0.6	a	9±0.4	a	0.4±0	a	9±0.6	a	6±0.4	a	0.4±0	a
IS7861	19±1.0	a	15±0.8	a	1±0.5	a	10±0.5	a	7±0.4	a	0.1±0	a
IS15148	36±1.2	c	33±1.9	a	6±0.5	b	17±1.0	b	12±1.0	a	1±0.3	a
IS16396	35±2.7	c	32±3.5	a	5±0.6	b	17±1.4	b	10±0.9	a	0.3±0	a
IS21622	32±2.3	c	25±1.5	a	0.7±0.1	a	22±1.6	c	16±1.3	b	0.3±0.2	a
IS23644	26±0.7	b	18±0.7	a	0.8±0.3	a	16±0.6	b	11±0.4	a	0.4±0	a
IS23645	25±0.9	b	21±0.9	a	1±0.2	a	15±0.8	b	12±0.7	a	0.5±0.1	a
IS23666	30±1.2	c	22±1.6	a	1±0.4	a	17±0.7	b	12±0.4	a	0.3±0.2	a
IS23669	24±1.5	b	17±1.1	a	0.5±0.2	a	14±0.8	b	10±1.2	a	0.5±0.3	a
IS23777	34±2.1	c	20±1.2	a	0.2±0	a	23±1.4	c	14±0.8	b	0.3±0.1	a
IS24072	25±2.5	b	18±2.1	a	1.3±0.4	a	17±0.8	b	11±0.6	a	0.2±0.1	a
IS24139	40±3.3	c	25±2.1	a	0.28±0	a	27±1.1	c	16±0.7	b	0.1±0	a
IS25077	38±1.8	c	22±1.4	a	0.5±0.3	a	22±1.8	c	12±0.6	a	0.2±0.2	a
IS25596	27±1.7	b	18±0.7	a	0.7±0.5	a	17±0.7	b	11±0.5	a	0.2±0	a
IS26457	36±2.9	c	27±2.8	a	1.5±0	a	19±1.4	c	13±0.9	a	0.5±0	a
IS26554	37±3.2	c	28±4.0	a	2±0.2	a	19±0.5	c	12±0.6	a	0.3±0	a
IS27490	26±1.5	b	16±0.9	a	0.3±0.1	a	18±1.3	c	11±0.8	a	0.1±0.1	a
IS29691	43±2.4	c	29±1.6	a	0.2±0	a	25±1.5	c	17±1.0	b	0.2±0	a
SSM215	17±0.6	a	11±0.4	a	0.3±0	a	10±0.5	a	7±0.4	a	0.3±0.2	a
SSM1057	19±0.6	a	13±0.3	a	0.6±0	a	12±0.7	a	8±0.4	a	0.5±0.2	a
CV(%)	11		24		57		16		23		71	

Table S3: Allele normalization procedure imposed on SSR allele sizes to genetically merge the GCP and the Embrapa collections (three SSR marker loci are shown for exemplification purposes). For control lines, the first column of each marker loci (gpsb067, mSbCIR246 and Xtxp114) shows allele sizes obtained at the Embrapa laboratory and the second one are the allele sizes obtained from the sorghum SSR kit (http://sat.cirad.fr/sat/sorghum_SSR_kit). Differences between these two values were used as a normalization constant. For the Embrapa collection, the first column of each marker loci are shows the allele sizes obtained at the Embrapa laboratory and the second one contain the allele sizes after normalization.

gpsb067			mSbCIR246		Xtxp114	
Correction	No correction		+2pb		-19pb	
Control lines						
	Embrapa	CGP	Embrapa	CGP	Embrapa	CGP
IS929	171.6	172	98.43	100	232.89	214
IS2156	169.52	170	98.52	100	229.88	211
IS2087	179.98	180	98.52	100	232.82	214
IS7889	171.6	172	98.43	100	0	214
IS11119	171.57	172	98.42	100	235.8	217
IS12531	175.76	176	98.51	100	232.85	214
SSM275	0	170	96.36	98	229.89	211
SSM379	179.97	180	98.5	100	232.82	214
SSM549	171.57	172:178	98.51	100	232.91	214
SSM1284	173.67	174	98.3	100	229.82	211
Embrapa collection						
	No normalized	Normalized	No normalized	Normalized	No normalized	Normalized
SC122-14	171.57	172	98.42	100	232.84	214
SC566-14	169.54	170	98.44	100	232.8	214
SC414-12	184.12	184	98.45	100	232.85	214
BR001B	179.97	180	98.5	100	232.81	214
5DX61-6-2	180.05	180	98.49	100	232.86	214
IS8577	171.65	172	98.51	100	232.91	214
CMSXS205	179.85	180	98.46	100	232.88	214
TX623B	179.84	180	98.45	100	232.86	214
IS3620C	0	0	96.39	98	229.88	211
V20-1-1	171.57	172	98.53	100	232.8	214
9910032	0	0	98.51	100	232.76	214
9DX9-11	179.86	180	98.5	100	232.89	214
IS10662B	179.9	180	98.58	100	232.82	214
9929034	171.56	172	93.87	96	232.73	214
TX2785N	179.94	180	98.51	100	232.89	214
SC283	180.07	180	98.43	100	0	0
BRANDES	179.87	180	98.43	100	232.87	214
SC175-14	179.88	180	98.52	100	232.87	214
TX2784N	171.56	172	98.38	100	232.84	214
CMSXS153	179.87	180	98.43	100	232.83	214
BR007B	179.92	180	98.5	100	232.89	214
BR012R	179.93	180	98.59	100	232.86	214

9929020	179.93	180	98.46	100	232.87	214
TX644B	171.53	172	98.53	100	229.84	211
TX430	179.82	180	98.38	100	232.77	214
QL3	179.91	180	98.54	100	232.74	214
SC549	179.87	180	98.52	100	232.74	214
CMSXS226R	179.83	180	0	0	232.8	214
TX643B	179.93	180	98.52	100	232.75	214
3DX-57-1-1-9-D	179.85	180	98.61	100	232.81	214
CMSXS116	179.99	180	98.59	100	232.81	214
CMSXS225R	179.9	180	98.5	100	0	0
ARG1B	179.85	180	98.5	100	232.86	214
TX642B	171.55	172	98.52	100	229.83	211
9929030	179.91	180	98.59	100	232.86	214
IS10317	179.92	180	98.51	100	232.87	214
ATF14	179.93	180	98.58	100	232.87	214
CMSXS227	180.05	180	98.59	100	232.88	214
CMSXS182	179.88	180	98.53	100	232.86	214
SC748	179.79	180	98.53	100	229.83	211
CMSXS156	179.87	180	98.52	100	232.78	214
CMSXS912	0	0	88.03	90	229.79	211
CMSXS180	179.88	180	98.52	100	232.75	214
CMSXS235	171.51	172	98.51	100	232.79	214
CMSXS230	0	0	0	0	232.75	214

Table S4: Details for SSR markers used in this study

Locus name	Chromosome	Repeat Motif	Number of alleles	PIC	Sequence of forward primer (5' to 3')	Sequence of reverse primer (5' to 3')
mSbCIR286	1	(AC)9	12	0.77	HEX-GCTTCTATACTCCCTCCAC	TTTATGGTAGGATGCTCTGC
mSbCIR306	1	(GT)7	3	0.58	6FAM-ATACTCTCGTACTCGGCTCA	GCCACTCTTTACTTTTCTTCTG
Xtxp320	1	(AAG)20	14	0.92	6FAM-TAAACTAGACCATATACTGCCATGATAA	GTGCAAATAAGGGCTAGAGTGTT
Xisep0310	2	(CCAAT)4	10	0.3	6FAM-TGCCTTGTGCCTTGTCTTATCT	GGATCGATGCCTATCTCGTC
Xgap84	2	(AG)14	22	0.82	HEX-CGCTCTCGGGATGAATGA	TAACGGACCACTAACAAATGATT
Xcup63	2	(GGATGC)4	5	0.34	6FAM-GTAAAGGGCAAGGCAACAAG	GCCCTACAAAATCTGCAAGC
mSbCIR238	2	(AC)26	23	0.89	NED-AGAAGAAAAGGGTAAGAGC	CGAGAAACAATTACATGAACC
mSbCIR223	2	(AC)6	7	0.67	6FAM-CGTTCCAATGACTTTTCTTC	GCCAATGTGGTGTGATAAAT
Xtxp114	3	(AGG)8	9	0.48	6FAM-CGTCTTCTACCGCGTCCT	CATAATCCCACTCAACAATCC
Xcup14	3	(AG)10	14	0.55	HEX-TACATCACAGCAGGGACAGG	CTGGAAAGCCGAGCAGTATG
Xcup61	3	(CAG)7	3	0.38	HEX-TTAGCATGTCCACCACAACC	AAAGCAACTCGTCTGATCCC
mSbCIR276	3	(AC)9	5	0.5	HEX-CCCCAATCTAACTATTTGGT	GAGGCTGAGATGCTCTGT
Xtxp021	4	(AG)18	19	0.68	NED-GAGCTGCCATAGATTGGTCG	ACCTCGTCCCACCTTTGTTG
Xtxp012	4	(CT)22	22	0.92	NED-AGATCTGGCGGCAACG	AGTCACCCATCGATCATC
mSbCIR246	5	(CA)7.5	6	0.19	HEX-TTTTGTGCACTTTTGAGC	GATGATAGCGACCACAAATC
SbAGB02	5	(AG)35	22	0.69	6FAM-CTCTGATATGTCGTTGTGCT	ATAGAGAGGATAGCTTATAGCTCA
Xtxp278	5	(TTG)12	12	0.39	6FAM-GGGTTTCAACTCTAGCCTACCGAACTTCCT	ATGCCTCATCATGGTTCGTTTGTCTT
Xtxp136	5	(GCA)5	2	0.32	6FAM-GCGAATAGCATCTTACAACA	ACTGATCATTGGCAGGAC
mSbCIR300	5	(GT)9	8	0.61	6FAM-TTGAGAGCGGCGAGGTAA	AAAAGCCCAAGTCTCAGTGCTA
Xtxp015	5	(TC)16	16	0.83	HEX-CACAAACACTAGTGCCTTATC	CATAGACACCTAGGCCATC
Xtxp057	6	(GT)21	20	0.81	NED-GGAACCTTTTGACGGGTAGTGC	CGATCGTGATGTCCCAATC
Xtxp265	6	(GAA)19	22	0.92	6FAM-GTCTACAGGCGTGCAAATAAAA	TTACCATGCTACCCCTAAAAGTGG
Xgap206	6	(AC)13/(AG)20	29	0.93	NED-ATTCATCATCCTCATCCTCGTAGAA	AAAAACCAACCCGACCCACTC
Xgap72	6	(AG)16	15	0.73	HEX-TGCCACCACTCTGGAAGGCTA	CTGAGGACTGCCCCAAATGTAGG
Xtxp040	7	(GGA)7	8	0.31	NED-CAGCAACTTGCACCTGTC	GGGAGCAATTTGGCACTAG
mSbCIR283	7	(CT)8 (GT)8.5	18	0.8	NED-TCCCTTCTGAGCTTGTAAT	CAAGTCACTACCAATGCAC
Xtxp141	7	(GA)23	17	0.87	HEX-TGTATGGCCTAGCTTATCT	CAACAAGCCAACCTAAA
mSbCIR262	7	(CATG)3.25	7	0.66	HEX-GCACCAAAATCAGCGTCT	CCATTTACCCGTGGATTAGT
Xtxp321	8	(GT)4+(AT)6+(CT)21	27	0.96	NED-TAACCCAAGCCTGAGCATAAGA	CCCATTACACATGAGACGAG
Xtxp273	8	(TTG)20	14	0.7	HEX-GTACCCATTTAAATTGTTTGCAGTAG	CAGAGGAGGAGGAAGAGAAGG
mSbCIR240	8	(TG)9	22	0.59	HEX-GTTCTTGCCCTACTGAAT	TCACCTGTAACCCTGTCTTC

gpsb123	8	(CA)7+(GA)5	8	0.66	6FAM-ATAGATGTTGACGAAGCA	GTGGTATGGGACTGGA
gpsb067	8	(GT)10	9	0.64	HEX-TAGTCCATACACCTTTCA	TCTCTCACACACATTCTTC
Xtxp145	9	(AG)22	21	0.88	NED-GTTCCTCCTGCCATTACT	CTTCCGCACATCCAC
Xcup02	9	(GCA)6	7	0.6	6FAM-GACGCAGCTTTGCTCCTATC	GTCCAACCAACCCACGTATC
Xtxp010	9	(CT)14	10	0.74	6FAM-ATACTATCAAGAGGGGAGC	AGTACTAGCCACACGTCAC
mSbCIR248	10	(GT)7.5	6	0.66	6FAM-GTTGGTCAGTGGTGGATAAA	ACTCCCATGTGCTGAATCT
mSbCIR329	10	(AC)8.5	7	0.72	NED-GCAGAACATCACTCAAAGAA	TACCTAAGGCAGGGATTG

Table S5: Membership assignment matrix to six subpopulations

Accession Name	Origin Country	ICRISAT	CIRAD		Q1	Q2	Q3	Q4	Q5	Q6	Clusters*	Linkage analysis
Q1 = guinea accessions from western Africa and guinea margaritifera												
SSM 275	BFA		Gm	Q1	0.99	0	0	0	0	0	2	
IS 23644	GMB	G	Gm	Q1	0.99	0	0	0	0	0	2	ns
SSM 505	NER		Gm	Q1	0.99	0	0	0	0	0	2	
IS 19953	SEN	G	Gm	Q1	0.99	0	0	0	0	0	2	
IS 27490	BFA	G	G	Q1	0.98	0	0	0	0.01	0	1	ns
SSM 249	BFA		G	Q1	0.98	0	0	0.01	0.01	0.01	1	
IS 7861	NGA	G	G	Q1	0.98	0.01	0.01	0.01	0.01	0	1	
IS 20097	SEN	G	G	Q1	0.98	0.01	0.01	0	0.01	0	1	
IS 6828	BFA	G	G	Q1	0.98	0.01	0	0	0.01	0.01	1	
IS 6745	BFA	G	G	Q1	0.98	0.01	0	0	0.01	0	1	
SSM 276	BFA		Gm	Q1	0.97	0	0	0	0.02	0	2	
IS 23666	GMB	G	Gm	Q1	0.97	0.01	0.01	0	0	0.01	2	s
IS 26110	MLI	G	G	Q1	0.97	0.01	0.01	0.01	0.01	0	1	
IS 9597	NER	G	Gm	Q1	0.97	0	0	0.01	0.01	0	2	
IS 26554	BEN	G	G	Q1	0.97	0.01	0.01	0.01	0	0.01	1	s
IS 20016	SEN	G	G	Q1	0.97	0.01	0.01	0.01	0.01	0.01	1	
IS 26457	BEN	G	G	Q1	0.97	0	0.01	0	0.01	0.01	1	ns
IS 20064	SEN	G	Gm	Q1	0.97	0.01	0	0	0.01	0.01	2	
IS 23669	GMB	G	Gm	Q1	0.97	0.01	0.01	0	0.01	0.01	2	
IS 23645	GMB	G	Gm	Q1	0.97	0	0	0	0.01	0.02	2	s
IS 3905	MLI	G	G	Q1	0.96	0.01	0.01	0	0.01	0.01	1	
IS 21849	SLE	G	Gm	Q1	0.96	0.01	0.02	0	0.01	0	2	
IS 27390	BFA	G	G	Q1	0.96	0.01	0	0.01	0.01	0	1	
SSM 1057	SEN		Gm	Q1	0.96	0	0	0.02	0	0.01	2	ns
SSM 625	NGA		DC	Q1	0.96	0.01	0.01	0.01	0.01	0.01		
IS3620C	NGA	G		Q1	0.95	0.01	0.01	0.02	0.01	0.01		
IS 26041	MLI	G	G	Q1	0.94	0.01	0	0	0	0.04	1	
IS 24887	NGA	G	G	Q1	0.94	0.01	0	0.01	0.03	0.01	1	
IS 17658	GHA	G	G	Q1	0.94	0.01	0.01	0.02	0.01	0.01	1	
IS 2430	NGA	G	Gm	Q1	0.93	0.03	0	0.01	0.02	0.01	2	
IS 10194	BFA	GB	B	Q1	0.92	0.01	0.02	0.01	0.03	0.02		

IS 25077	GHA	G	G	Q1	0.92	0.01	0.01	0.01	0.01	0.05	1	
SSM 205	BFA		G	Q1	0.91	0.01	0.01	0.05	0.01	0.01	1	
IS 16396	CMR	G	G	Q1	0.9	0.02	0.07	0	0	0.01	1	s
IS 25733	MLI	G	G	Q1	0.89	0.05	0.01	0.01	0.01	0.04	1	
SSM 1046	SEN		G	Q1	0.88	0.02	0.01	0	0.09	0	1	
SSM 232	BFA		G	Q1	0.81	0.07	0.01	0.01	0.03	0.07		
SSM 964	SEN		D	Q1	0.75	0.13	0.03	0.02	0.05	0.04	unc.	
IS 10844	TCD	G	GC	Q1	0.73	0.1	0.06	0.01	0.05	0.05	1	
IS 7889	NGA	G	G	Q1	0.72	0.06	0.01	0.04	0.06	0.12	1	
SC566-14	NGA	C		Q1	0.68	0.01	0.28	0.01	0.01	0.01		s
SC549	NGA	G		Q1	0.64	0.01	0.33	0.01	0.01	0		s
SSM 501	NER		DC	Q1	0.58	0.01	0.01	0.01	0.4	0	unc.	
IS 25702	MLI	GC	GC	Q1	0.52	0.42	0.03	0.01	0.01	0.01	unc.	
SSM 557	NER		G	Q1	0.47	0.01	0.01	0.01	0.03	0.48	unc.	
IS 19466	LKA	G	G	Q1	0.46	0.04	0.08	0.41	0	0.01	unc.	
IS 25499	BDI	C	C	Q1	0.42	0.4	0.04	0.01	0.03	0.11	10	
Q2 = caudatum accessions from africa and group of transplanted caudatum and durra accessions from lake chad region												
V20-1-1	UGA			Q2	0	0.98	0.01	0	0	0		
3DX-57-1-1-9-D				Q2	0	0.98	0.01	0.01	0	0		s
IS 13452	ZWE	GC	C	Q2	0	0.98	0.01	0.01	0.01	0	5	
SC112-14	ETH	C		Q2	0.01	0.97	0.01	0	0.01	0		ns
IS 2807	ZWE	C	C	Q2	0	0.97	0.01	0	0	0	5	
SSM 1611	TCD		D	Q2	0	0.97	0	0	0.01	0.01	6	
SSM 1592	TCD		D	Q2	0	0.97	0	0	0	0.01	6	ns
5DX61-6-2	UGA			Q2	0	0.97	0.01	0.01	0	0.01		
IS 16044	CMR	C	C	Q2	0.01	0.97	0.01	0.01	0	0.01	5	
SSM 552	NER		C	Q2	0	0.97	0.01	0	0	0.01	5	
SSM 1123	NER		C	Q2	0	0.97	0.01	0.01	0	0	5	
IS8577				Q2	0.01	0.97	0.01	0	0	0.01		s
IS 2814	ZWE	C	C	Q2	0.01	0.97	0	0.01	0	0.02	5	
IS 13848	ZAF	C	C	Q2	0.01	0.96	0.01	0.01	0.01	0	5	
IS 2787	KEN	C	C	Q2	0	0.96	0.01	0.01	0.01	0.01	5	
SSM 1267	CMR		DC	Q2	0	0.96	0.02	0.01	0.01	0.01	5	
IS 3073	SDN	C	C	Q2	0.01	0.96	0.01	0	0.01	0.01	5	
SSM 19	CMR		D	Q2	0.01	0.96	0.01	0	0.01	0.01	6	

IS 33116	CMR	C	C	Q2	0.01	0.96	0.01	0.01	0.01	0.02	6	ns
IS 3511	SDN	KC	C	Q2	0.01	0.95	0.01	0.01	0.03	0		
IS 2848	ZAF	C	C	Q2	0.01	0.95	0.01	0.03	0.01	0	5	
IS 15148	CMR	C	C	Q2	0.01	0.95	0.02	0.01	0.01	0.01	5	
IS 10234	CAF	GC	C	Q2	0.01	0.95	0.01	0.03	0.01	0.01	5	
IS 16173	CMR	C	C	Q2	0.01	0.94	0.01	0.01	0	0.04	5	
IS 29876	ZWE	GC	C	Q2	0.01	0.94	0.01	0.03	0.01	0.01	5	
IS 16545	CMR	C	C	Q2	0	0.94	0.01	0.02	0	0.03	5	
9929020	BRA			Q2	0.01	0.94	0.03	0.01	0	0.01		
IS 15752	CMR	C	C	Q2	0.02	0.94	0.01	0.02	0.01	0.01		
SSM 1102	TCD		DC	Q2	0	0.94	0.01	0.02	0.01	0.02	5	ns
SSM 12	CMR		D	Q2	0.01	0.93	0.01	0.01	0.04	0.01	6	
IS 15443	CMR	C	C	Q2	0.01	0.93	0.04	0.01	0.01	0.01	5	
IS 10876	NGA	GC	C	Q2	0.03	0.92	0.01	0.01	0.02	0.01	5	
IS 7287	NGA	C	G	Q2	0.01	0.92	0.02	0.01	0.01	0.04		
IS 2262	SDN	KC	C	Q2	0.01	0.92	0.03	0.02	0.02	0.01	5	
CMSXS235	BRA			Q2	0	0.91	0.01	0.06	0	0.01		
9DX9-11	UGA			Q2	0	0.9	0.05	0.03	0.01	0		
SSM 1103	TCD		DC	Q2	0.03	0.89	0.04	0.01	0.02	0.01	6	
IS 26833	SDN	C	C	Q2	0.02	0.89	0.02	0.05	0.01	0.01	5	
IS 16186	CMR	C	D	Q2	0.06	0.87	0.01	0	0.05	0	6	s
IS 9331	ZAF	K	C	Q2	0.03	0.83	0.1	0.02	0.01	0.01		
SSM 547	NER		C	Q2	0.11	0.82	0.03	0.02	0.02	0.01	6	
IS 33261	CMR	C	C	Q2	0.02	0.78	0.01	0.03	0.14	0.02	6	
IS 16125	CMR	C	C	Q2	0.04	0.78	0.12	0.01	0.04	0.01	6	
SSM 29	CMR		D	Q2	0.02	0.72	0.01	0.01	0.23	0.02	6	
IS 33353	KEN	C	C	Q2	0.02	0.7	0.04	0.01	0.02	0.21	10	
IS 10882	NGA	GC	C	Q2	0.08	0.66	0.16	0.01	0.09	0.01	unc.	
IS 8685	SDN	C	C	Q2	0.03	0.63	0.02	0.01	0.1	0.21	unc.	
IS 303	CHN	KB	B	Q2	0.01	0.62	0.01	0.02	0.31	0.04	unc.	
SC114-12	SDN	KC		Q2	0.06	0.6	0.08	0.1	0.13	0.04		unc.
IS 19026	SDN	C	DC	Q2	0.03	0.57	0.21	0.17	0.02	0	unc.	
IS 21124	KEN	C	C	Q2	0.01	0.5	0.05	0.01	0.41	0.02	5	
IS 8882	UGA	C	C	Q2	0.1	0.47	0.08	0.05	0.19	0.12	10	
SC748	SDN	GC		Q2	0	0.45	0.52	0.02	0.01	0.01		

9929034				Q2	0.01	0.45	0.32	0.21	0.01	0.01		
Brandes	USA			Q2	0.01	0.43	0.34	0.18	0.03	0.01		
IS 25596	RWA	C	C	Q2	0.01	0.4	0.03	0.01	0.18	0.37	10	
IS 16101	CMR	D	D	Q2	0.3	0.36	0.01	0	0.33	0	unc.	
Q3 = lines from brazil and usa												
CMSXS116B	ETH	C		Q3	0	0.01	0.98	0	0	0		
CMSXS182	BRA			Q3	0	0.01	0.98	0	0	0		
BTX643	USA			Q3	0	0	0.97	0.01	0.01	0		
9910032	BRA			Q3	0.01	0.01	0.97	0.01	0.01	0.01		
BTX623	USA			Q3	0	0.01	0.97	0.02	0	0		
TX430	USA			Q3	0.02	0.01	0.94	0.01	0.01	0.02		
CMSXS180	BRA			Q3	0	0	0.93	0.06	0.01	0		
BTX644	USA			Q3	0.01	0.01	0.92	0	0.06	0.01		
SC175-14	ETH	C		Q3	0.01	0.06	0.92	0	0.01	0		s
CMSXS227	BRA			Q3	0	0.04	0.92	0.02	0	0.02		
CMSXS226R	BRA			Q3	0	0.05	0.87	0.04	0	0.03		
CMSXS156	USA			Q3	0	0	0.83	0.15	0	0.01		
BR012R	BRA			Q3	0	0.15	0.83	0.01	0	0.01		s
IS 21891	USA	C	C	Q3	0.19	0.04	0.76	0.01	0.01	0.01	5	
CMSXS225R				Q3	0	0.01	0.71	0.17	0	0.12		s
ARG1	USA/ARG			Q3	0	0.02	0.66	0.32	0	0		
IS 19132	SDN	GC	CB	Q3	0.01	0.27	0.64	0.06	0.02	0.01	unc.	
IS 19453	BWA	D	D	Q3	0.02	0.01	0.63	0.01	0.32	0	3	
9929030	BRA			Q3	0.08	0.23	0.61	0.06	0.02	0		
BTX642	ETH	D		Q3	0.02	0.01	0.53	0.01	0.43	0.01		
TX2785N	USA			Q3	0.01	0.01	0.52	0.01	0.45	0		
SC283	TZA	G		Q3	0	0.02	0.51	0.13	0	0.33		s
ATF14	BRA			Q3	0	0.02	0.5	0.44	0	0.03		
SSM 215	ETH		C	Q2	0.02	0.44	0.48	0.03	0.02	0.01	5	ns
Q4 = kafir accessions from southern africa												
IS10317				Q4	0	0	0.01	0.98	0	0		
IS 29496	LSO	K	K	Q4	0	0.01	0	0.98	0	0	7	
IS 22334	BWA	K	K	Q4	0	0	0	0.98	0	0.01	7	
IS 30030	ZWE	K	K	Q4	0.01	0	0	0.98	0	0.01		
IS 29233	SWZ	K	K	Q4	0	0.01	0	0.98	0.01	0	7	

IS 29409	LSO	K	K	Q4	0	0.01	0	0.98	0.01	0	7	
IS 29606	ZAF	K	K	Q4	0	0.01	0.01	0.98	0	0.01	7	
IS 29407	LSO	K	KC	Q4	0.01	0	0	0.98	0.01	0.01	7	
IS 9527	ZAF	K	K	Q4	0	0.01	0.01	0.98	0	0.01	7	
IS 29911	ZWE	C	KC	Q4	0	0	0	0.97	0.01	0.01		
IS 29310	SWZ	GC	KC	Q4	0	0.01	0.01	0.97	0	0.01	7	
IS 13989	ZAF	K	K	Q4	0	0.01	0.01	0.97	0	0.01	7	
IS 29872	ZWE	C	KC	Q4	0	0.01	0.01	0.97	0.01	0.01	7	
IS 29226	SWZ	GC	KB	Q4	0	0	0.01	0.96	0	0.02		
IS 13791	ZAF	K	KC	Q4	0	0	0.01	0.96	0.02	0	7	
IS 2398	ZAF	K	K	Q4	0	0.01	0.01	0.96	0	0.01	7	
CMSXS153	UGA			Q4	0	0	0.04	0.95	0	0		
IS 9468	ZAF	K	K	Q4	0	0.01	0	0.95	0.01	0.03	7	
IS 27164	ZWE	K	K	Q4	0	0.01	0.01	0.95	0.02	0.02	7	
IS 22330	BWA	D	D	Q4	0.01	0	0	0.95	0.02	0.02	7	
QL3	AUS			Q4	0	0	0.05	0.94	0	0		
IS 22294	BWA	K	K	Q4	0.01	0.01	0.03	0.92	0.01	0.02		
IS 22239	BWA	K	K	Q4	0.01	0.01	0.01	0.91	0.03	0.03		
IS 9303	ZAF	K	K	Q4	0	0.01	0.02	0.91	0.05	0.01	7	
IS 27891	ZAF	GC	GC	Q4	0.01	0.01	0	0.91	0.04	0.04	7	
IS 13845	ZAF	K	K	Q4	0	0	0	0.89	0	0.09	7	
IS 29375	LSO	GC	GC	Q4	0.01	0.01	0.07	0.88	0.02	0.01	7	
IS 22287	BWA	K	K	Q4	0.04	0.01	0.02	0.87	0.03	0.03	7	
IS 29569	LSO	K	K	Q4	0.01	0.01	0	0.87	0.04	0.07	7	
IS 24481	ZAF	K	K	Q4	0	0.09	0	0.83	0.07	0.01	7	
IS 13926	ZAF	K	KC	Q4	0	0	0	0.82	0.01	0.16	7	
BR007B	USA			Q4	0	0.01	0.24	0.73	0.01	0		
IS10622B				Q4	0	0.01	0.32	0.65	0.01	0.01		
IS 22332	BWA		KC	Q4	0.01	0.01	0.01	0.63	0.34	0.01	7	
IS 22282	BWA	B	B	Q4	0	0.01	0	0.6	0.01	0.38	7	
IS 24139	TZA	G	G	Q4	0.02	0.01	0	0.58	0.01	0.39	unc.	ns
BR001B	USA			Q4	0.02	0	0.39	0.58	0.01	0		
BTX631	USA	KC		Q4	0	0.01	0.42	0.56	0	0.01		
IS 26731	ZAF	B	B	Q4	0	0.25	0.02	0.54	0.01	0.18	7	
IS 30175	ZWE	G	G	Q4	0	0.02	0	0.47	0	0.5	8	

CMSXS230				Q4	0.07	0.36	0.05	0.44	0.06	0.02	
IS 24009	USA	C	C	Q4	0.01	0.26	0.27	0.38	0.02	0.07	unc.
IS 12447	SDN	DC	GC	Q4	0.26	0.1	0.24	0.36	0.04	0.01	unc.
Q5 = durra, bicolor and caudatum accessions from eastern africa and asia											
IS 30436	CHN	C	C	Q5	0	0	0	0	0.98	0	4
IS 929	SDN	D	D	Q5	0	0.01	0	0	0.98	0	3
SSM 1049	SEN		B	Q5	0.01	0	0	0.01	0.98	0	unc.
IS 30451	CHN	C	C	Q5	0	0	0.01	0	0.98	0.01	4
IS 30443	CHN	C	CB	Q5	0	0.01	0	0.01	0.98	0	4
IS 4285	IND	D	D	Q5	0	0	0.01	0	0.98	0	3
IS 6193	IND	D	D	Q5	0	0.01	0	0	0.98	0.01	3
IS 30538	KOR	B	B	Q5	0	0	0	0.01	0.98	0.01	4
IS 30405	CHN	DC	BC	Q5	0	0.01	0	0.01	0.98	0.01	4
SSM 1284	ZAR		B	Q5	0.01	0	0.01	0	0.98	0.01	3
IS 5867	IND	DC	DC	Q5	0.01	0.01	0.01	0.01	0.97	0	3
IS 6118	IND	DC	DC	Q5	0.01	0.01	0	0	0.97	0.01	3
IS 30352	CHN	DC	CB	Q5	0.01	0.01	0	0.01	0.97	0.01	4
IS 3971	IND	D	D	Q5	0	0.01	0.01	0.01	0.97	0	3
IS 6351	IND	D	D	Q5	0	0	0.01	0.01	0.97	0	3
IS 30441	CHN	C	C	Q5	0	0.01	0.01	0.01	0.97	0	4
IS 2263	SDN	D	D	Q5	0.01	0.01	0	0	0.97	0.01	3
IS 30417	CHN	DC	BC	Q5	0.01	0.01	0	0.01	0.97	0.01	4
IS 11026	ETH	D	D	Q5	0.01	0.01	0.01	0.01	0.97	0	3
IS 28409	YEM	D	D	Q5	0.01	0.01	0.01	0	0.97	0	3
IS 20351	NER	D	D	Q5	0.01	0.01	0.01	0.01	0.97	0	3
IS 5972	IND	D	D	Q5	0	0.01	0.01	0	0.97	0.01	3
IS 3771	CHN	C	CB	Q5	0.01	0.02	0	0.01	0.97	0	
IS 4821	IND	D	D	Q5	0.01	0.01	0.01	0.01	0.96	0.01	3
IS 5430	IND	B	B	Q5	0.01	0.01	0	0.01	0.96	0.02	
IS 1398	SDN	D	D	Q5	0.02	0.01	0	0	0.96	0.01	3
IS 19847	IND	DC	DC	Q5	0.01	0.01	0.01	0.01	0.95	0.01	3
IS 32569	SOM	D	D	Q5	0.01	0.01	0.02	0.02	0.95	0	3
IS 28645	YEM	DC	DC	Q5	0.04	0	0.01	0	0.95	0	3
IS 12542	ETH	D	D	Q5	0.01	0.01	0.03	0.01	0.95	0	
IS 11827	ETH	D	D	Q5	0.02	0.02	0.01	0.01	0.94	0	3

SSM 546	NER		D	Q5	0.04	0.01	0.01	0.01	0.94	0	3
IS 4027	IND	B	B	Q5	0.01	0.01	0	0	0.93	0.04	3
IS 30385	CHN	C	C	Q5	0	0	0	0	0.93	0.05	4
IS 12804	TUR	B	B	Q5	0.01	0.01	0.01	0.02	0.93	0.03	unc.
IS 30400	CHN	C	KC	Q5	0.01	0	0	0.06	0.92	0.01	4
IS 20864	CHN	C	C	Q5	0	0.01	0.07	0.01	0.9	0.01	unc.
IS 31681	DZA	B	B	Q5	0.06	0.02	0	0	0.9	0.01	unc.
IS 13113	IND	B	B	Q5	0.01	0.03	0.03	0.02	0.9	0.01	unc.
IS 13	USA	B	B	Q5	0.01	0.01	0.01	0.07	0.9	0.01	unc.
IS 12179	ETH	B	B	Q5	0.02	0.01	0.01	0.03	0.9	0.03	unc.
SSM 1370	ZAF		B	Q5	0	0.03	0.03	0.04	0.86	0.04	unc.
IS 20727	USA	B	B	Q5	0	0.02	0.02	0.09	0.86	0.01	unc.
IS 12169	ETH	B	B	Q5	0.08	0.01	0.01	0.02	0.84	0.04	unc.
IS 20706	USA	B	B	Q5	0.05	0.07	0.02	0.01	0.83	0.02	3
IS 3780	CHN	GB	B	Q5	0.05	0.01	0.03	0.04	0.77	0.1	unc.
IS 20689	USA	B	B	Q5	0.02	0.05	0.11	0.02	0.77	0.04	unc.
IS24503	SEN			Q5	0.03	0.02	0.01	0.06	0.72	0.17	
TX2784	USA			Q5	0.01	0.07	0.07	0.16	0.7	0	
IS 12531	ETH	B	B	Q5	0.01	0.01	0.04	0.15	0.65	0.15	unc.
IS 2156	NGA	B	B	Q5	0.03	0.01	0.02	0.01	0.59	0.34	unc.
CMSXS912	USA			Q5	0.02	0.33	0.04	0.09	0.51	0.01	
SSM 973	SEN		D	Q5	0.01	0.01	0.47	0.01	0.5	0	3
IS 22893	SDN	GC	GC	Q5	0.08	0.2	0.12	0.04	0.29	0.27	10
Q6 = guinea accessions from southern africa and asia											
IS 4963	IND	G	G	Q6	0	0	0	0	0	0.99	9
IS 3958	NPL	G	G	Q6	0	0	0	0	0	0.99	9
IS 3959	NPL	G	G	Q6	0	0	0	0	0	0.99	9
IS 3967	IND	G	G	Q6	0	0	0	0	0	0.98	9
IS 3957	NPL	G	G	Q6	0	0.01	0	0	0	0.98	9
IS 14417	MWI	G	G	Q6	0	0.01	0.01	0	0.01	0.98	8
IS 14414	MWI	G	G	Q6	0	0.01	0.01	0.01	0.01	0.97	8
IS 14317	SWZ	G	G	Q6	0	0.01	0.01	0.04	0	0.94	8
SSM 379	MLI		G	Q6	0	0.01	0.01	0.06	0	0.91	8
IS 3421	IND	G	G	Q6	0	0.04	0.01	0.01	0.05	0.89	9
IS 23254	ZMB	B	B	Q6	0	0.01	0.03	0.05	0.03	0.88	8

IS 23777	MWI	G	G	Q6	0.01	0.01	0.02	0.08	0.04	0.85	8	
IS 19455	BWA	G	Gm	Q6	0	0.01	0.05	0.1	0.02	0.82	8	
IS 14331	ZAF	G	G	Q6	0.01	0	0.01	0.17	0	0.81	8	
IS 19685	IND	G	G	Q6	0	0	0.01	0	0.19	0.8	9	
IS 12931	CHN	G	G	Q6	0.01	0.01	0	0	0.21	0.77	9	
IS 29691	ZWE	G	G	Q6	0.01	0	0	0.21	0.01	0.77	8	ns
IS 31559	BDI	C	C	Q6	0.01	0.19	0.02	0.01	0.01	0.77	10	
IS 21519	MWI	G	G	Q6	0.01	0.01	0.05	0.02	0.19	0.72	8	ns
IS 24072	TZA	G	G	Q6	0.02	0.03	0.04	0.23	0.02	0.67	unc.	
IS 21622	MWI	G	G	Q6	0.01	0.01	0.01	0.34	0.01	0.63	8	ns
IS 14351	MWI	G	G	Q6	0	0.01	0	0.36	0	0.62	8	s
IS 27146	ZWE	D	D	Q6	0.01	0.01	0.14	0.28	0	0.56	unc.	
IS 23178	ZMB	B	B	Q6	0.09	0.01	0.02	0.03	0.32	0.54	unc.	
IS 23100	TZA	G	G	Q6	0	0.02	0.24	0.19	0.01	0.54	unc.	
IS 2416	ZAF	K	K	Q6	0.01	0.34	0.01	0.14	0.01	0.5	unc.	
IS 23142	TZA	D	D	Q6	0.39	0.01	0.01	0.08	0.06	0.46	unc.	s
IS 11119	ETH	C	C	Q6	0.11	0.29	0.08	0.01	0.06	0.46	10	
IS 10801	TCD	GC	GC	Q6	0.35	0.07	0.13	0.05	0.02	0.37	unc.	s

* Cluster classification according to Deu et al (2006); 1= guinea from western Africa; 2= guinea margaritifera subrace from western Africa; 3= durra accessions from central and eastern Africa and from Asia; 4= bicolor and caudatum accessions from China; 5= caudatum accessions from Africa; 6= group of transplanted caudatum and durra accessions from Lake Chad region; 7= accessions of kafir race from southern Africa; 8= guinea accessions from southern Africa; 9= guinea accessions from Asia; 10= accessions of the caudatum race from the African Great Lakes region

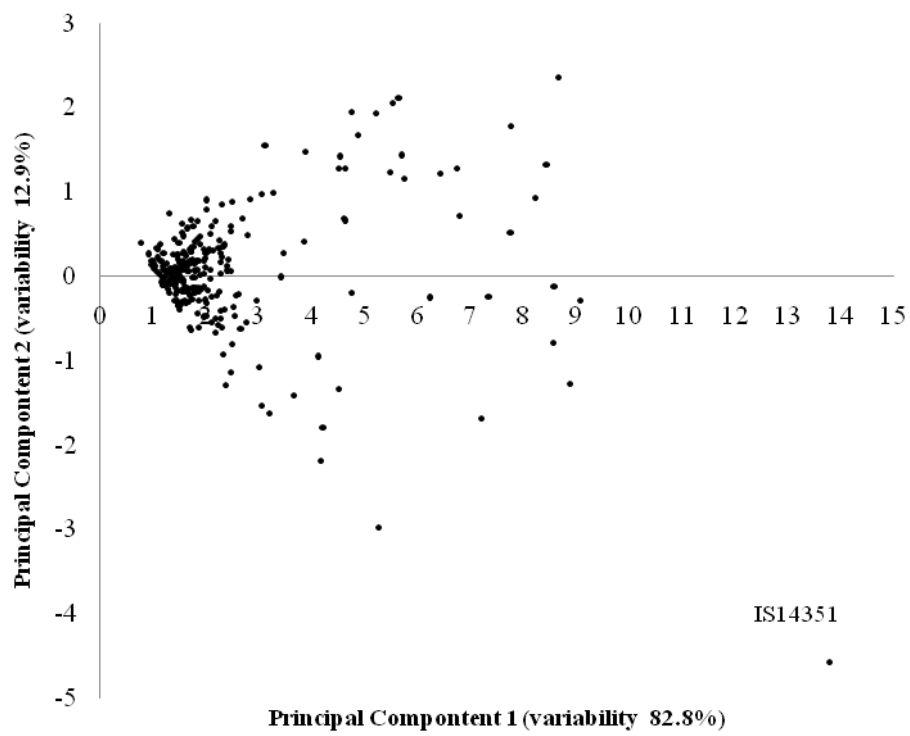


Figure S1: Principal Component Plot for $RNRG_{3d}$, $RNRG_{5d}$, VRD and b index (standardized variables). The cumulative proportion of total variance (%) explained are shown.

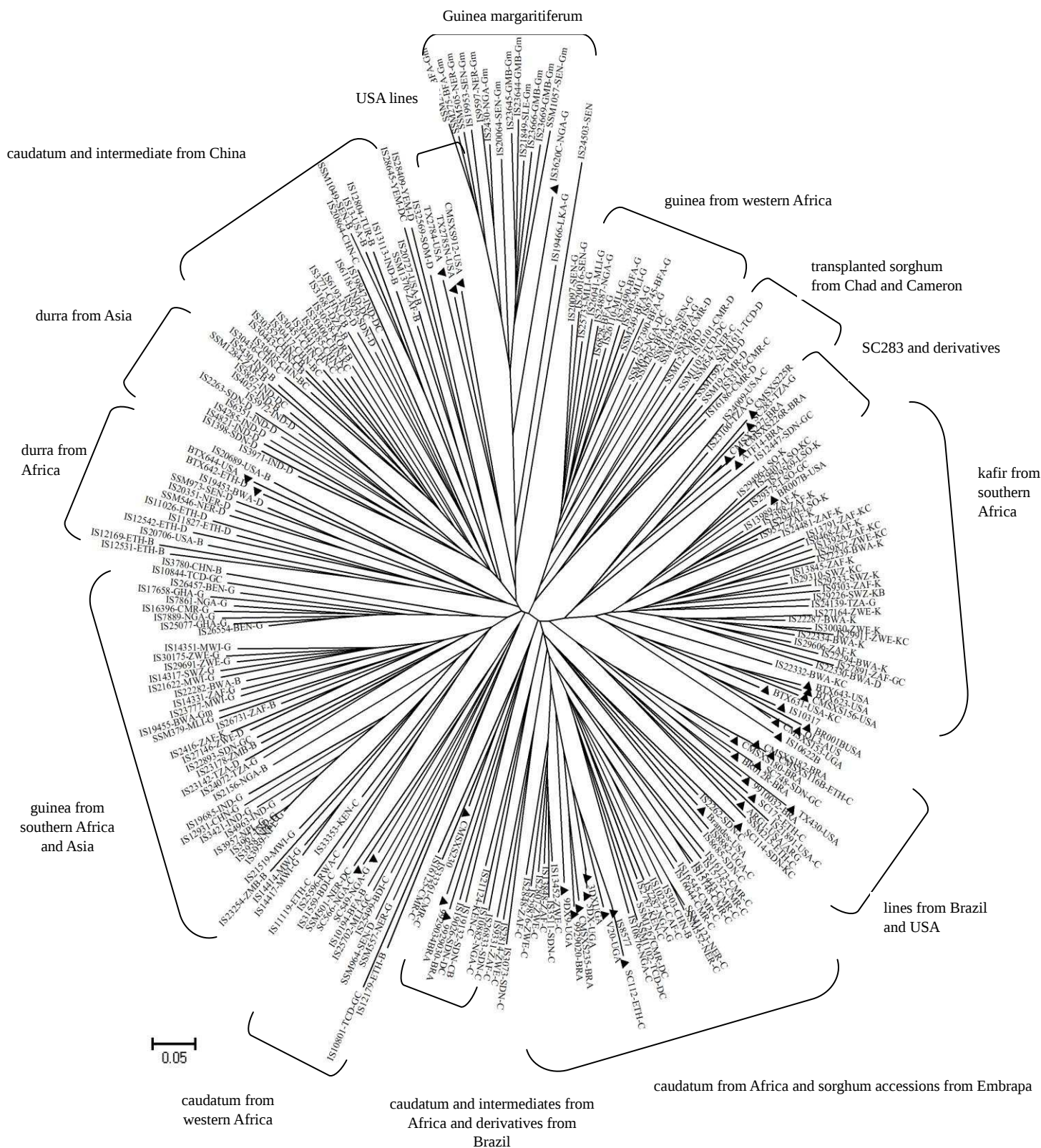


Figure S2: Cluster analysis using 254 sorghum accessions based on of a Nei and Li (1979) matrix of genetic dissimilarities and Neighbor-Joining procedure. Black triangle indicates lines from the Embrapa collection. Clusters cited on the text are identified.

Capítulo 2

Mapeamento Associativo para Tolerância ao Alumínio em Regiões Candidatas do Loco *Alt_{SB}* de Sorgo

1. Introdução

A toxidez causada pelo Al é um importante fator que limita o cultivo do sorgo e de outras espécies cultivadas em solos ácidos. Quando genótipos sensíveis de sorgo são cultivados em solos com $\text{pH} < 5.0$, observa-se uma inibição rápida do crescimento radicular, comprometendo a absorção de água e nutrientes pela planta (Kochian et al. 2004). O principal mecanismo fisiológico relacionado com a tolerância ao Al em plantas envolve a exudação de ácidos orgânicos (Kochian et al. 2004). Recentemente, um membro da família multigênica *Multidrug And Toxic compound Extrusion* (MATE), um transportador de citrato ativado por alumínio, foi clonado e validado como responsável pelo loco de tolerância ao Al, *Alt_{SB}* (Magalhaes et al. 2007), que responde por grande parte da variação fenotípica da tolerância ao Al em sorgo (Magalhaes et al. 2004; Caniato et al. 2007).

Variações fenotípicas pronunciadas para tolerância ao Al entre linhagens de sorgo são condicionadas por uma série alélica no loco *Alt_{SB}*, ainda que haja evidências indicando a ocorrência de outros genes de tolerância ao Al distintos do gene *Alt_{SB}* (Caniato et al. 2007). Ainda que a tolerância ao Al seja um fenômeno raro em sorgo (capítulo 1), uma frequência relativamente alta de acessos tolerantes foi relatada nas raças *caudatum* e *guinea* (Reddy et al. 2006). A recente avaliação de uma coleção nuclear de sorgo para tolerância ao Al revelou que entre os 16 acessos classificados como tolerantes (CLR $\geq 80\%$), sete eram integrantes da raça *guinea*, dois eram da raça *caudatum*, um da raça *durra* e seis não apresentavam classificação racial. Os acessos que apresentaram resposta intermediária de tolerância foram freqüentemente encontrados nas raças *caudatum* e *guinea*. Adicionalmente, as análises genéticas conduzidas em dezoito acessos tolerantes e de tolerância intermediária revelaram que a tolerância ao Al em seis deles é condicionada pelo loco *Alt_{SB}* (capítulo 1), apesar da funcionalidade do loco *Alt_{SB}* não ter sido descartada para os 12 restantes devido ao caráter recessivo da tolerância ao Al entre esses acessos.

Durante o trabalho de clonagem posicional do gene *Alt_{SB}*, Magalhães et al (2007) determinaram uma região de 24,6kpb (kilo pares de bases) no cromossomo 3 de sorgo na qual o loco de tolerância está localizado. A região flanqueada por marcadores moleculares

contém três quadros de leitura aberta (*open-reading frames*, ORFs) incluindo um homólogo de sorgo de um gene da família multigênica MATE, *SbMATE*, que foi posteriormente validado como responsável pelo loco *Alt_{SB}*. Toda a região determinada pelo mapeamento de alta resolução foi sequenciada nos parentais da população de mapeamento SC283 (tolerante ao Al) e BR007 (sensível ao Al). Foram encontrados os seguintes polimorfismos (Figura 1): 1) uma inserção de transposon do tipo *Miniature Inverted Repeat Transposable Element* (MITE, Bureau et al. 1992; Wessler et al. 1995), localizado na região promotora do gene *SbMATE*; 2) seis polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms*, SNPs) e um polimorfismo do tipo inserção-deleção (indel) no segundo intron do gene; 3) dois SNPs e um indel na região 3' do gene *Alt_{SB}*, após o códon de terminação e um indel de 19pb também na região 3'. Os polimorfismos localizados nessas regiões foram considerados candidatos como polimorfismos causativos de efeitos fenotípicos (*Quantitative Trait Nucleotide*, QTN) entre SC283 e BR007.

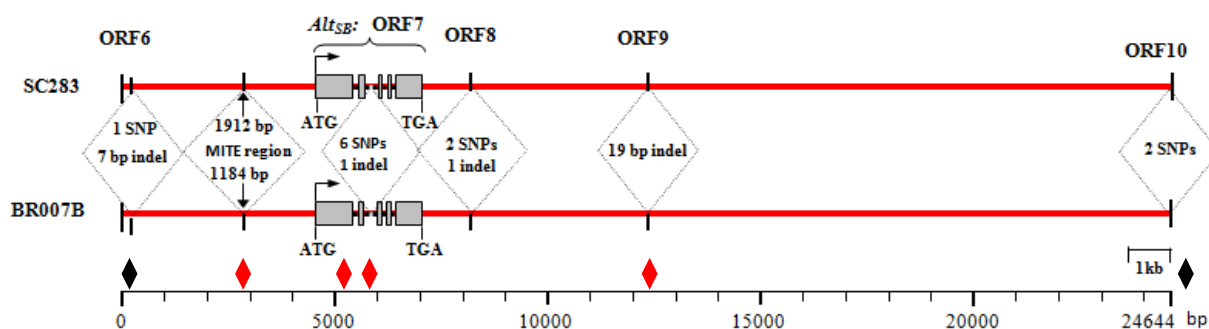


Figura 1. Seis regiões do loco *Alt_{SB}* amostradas para o mapeamento associativo. O gene *SbMATE* (ORF7, exons representados por caixas cinzas e introns por linhas pretas) bem como um gene similar a uma proteína hipotética (ORF8) e outro similar ao gene sacarose fosfato sintase (ORF9) foram flanqueados por um marcador correspondente a um indel de 7pb entre os parentais SC283 e BR007 na ORF6 e um SNP G/A na ORF10. Amplicons compreendendo os dois locos marcadores além de amplicons no primeiro exon (E1) e no segundo intron (I2) do gene *SbMATE*, na ORF9 e na ORF10 foram selecionados para as análises de associação. A região do MITE foi amplificada com primers flanqueando o indel e o loco foi genotipado para a análise associativa. As regiões estudadas na análise associativa estão marcadas com diamantes acima da linha de escala, onde diamantes vermelhos indicam regiões candidatas para QTNs e pretos indicam os dois locos flanqueadores.

Devido à alta correlação observada entre expressão gênica e tolerância ao Al ($r=0.98$, $p<0.001$) em 11 linhagens de sorgo (Magalhaes et al. 2007) e à natureza monomórfica da região codante do gene *Alt_{SB}*, fatores regulatórios fora da região codante do gene podem controlar níveis de transcritos do gene *Alt_{SB}* sendo, dessa maneira, responsáveis

pela variação alélica no loco. Sendo assim, a busca pelo(s) QTN(s) foi estendida para fora da região codante do gene *Alt_{SB}*. Foram identificados um SNP e um indel de 7pb após a extremidade 5' do gene *SbMATE* e dois SNPs na ORF10, após a extremidade 3' do gene de tolerância. Foram detectados um evento de recombinação entre cada um desses polimorfismos e o gene *SbMATE*, o que os descarta como QTNs.

No contexto desse capítulo, utilizando-se um painel de sorgo que captura grande parte da variabilidade genética da espécie (Deu et al. 2006) e outros 45 acessos do programa de melhoramento da Embrapa Milho e Sorgo enriquecidos para tolerância ao AI (Caniato et al. 2007), o mapeamento associativo foi usado visando à identificação de variantes responsáveis pelos diferentes alelos do loco *Alt_{SB}* e a identificação de QTNs ou locos em desequilíbrio de ligação com os mesmos, de maneira a elucidar a estrutura funcional do loco *Alt_{SB}*. Finalmente, a diversidade haplotípica do loco *Alt_{SB}* foi analisada para construir inferências sobre a origem dos haplótipos altamente tolerantes em um contexto filogeográfico, tendo como pano de fundo a dinâmica da domesticação da espécie e a importância da seleção na evolução do loco *Alt_{SB}*. Do ponto de vista de aplicação prática, foram identificados marcadores SNP que discriminam haplótipos-alvo no loco *Alt_{SB}* (*tag* SNPs), que podem ser utilizados para seleção assistida para tolerância ao AI em populações estruturadas e não estruturadas de sorgo.

2. Material e Métodos

2.1. Material Genético

O germoplasma avaliado consiste de 209 acessos mantidos pelo *Centre de Cooperation Internationale en Recherche Agronomique pour le Développement* (CIRAD - França) e pelo *International Crops Research Institute for the Semi-Arid Tropics* (ICRISAT, Patancheru - Índia). Essa coleção constitui uma amostra representativa da diversidade genética da espécie e foi elaborada com base em classificação racial, origem geográfica e resposta ao fotoperiodismo, bem como em padrões de diversidade avaliados com marcadores microssatélite (SSR). Esse germoplasma é comumente utilizado por agricultores na África e na Ásia e melhoristas do ICRISAT e CIRAD (Deu et al. 2006). Além disso, fizeram parte deste estudo outras 45 linhagens frequentemente usadas pelo programa de melhoramento da Embrapa Milho e Sorgo bem como por programas de melhoramento dos Estados Unidos. Nesse painel estão incluídas linhagens elite, landraces e derivativos de melhoramento, para os quais há disponibilidade de informações sobre a tolerância ao AI

(Caniato et al. 2007).

2.2. Variáveis Avaliadas

A tolerância ao Al em sorgo foi avaliada com base no crescimento radicular em solução nutritiva nas concentrações de 0 e 148 μM que correspondem às atividades livres de Al de {0} e {27} μM Al, a pH 4.0 (estimativas de atividade livre obtidas com o software de especiação GEOCHEM-PC, Parker et al. 1995). Após 4 dias de germinação, as plântulas foram transferidas para solução nutritiva completa sem Al e mantidas nesta condição por um período de 24h de aclimação quando comprimento radicular inicial foi obtido (C_{0d}). A solução nutritiva foi então trocada por solução contendo Al nas atividades de {0} e {27} μM Al. O comprimento radicular foi novamente obtido após 3 e 5 dias de crescimento em solução nutritiva contendo {0} e {27} μM Al^{3+} , (C_{3d} e C_{5d}). O crescimento radicular líquido (CL) após 3 e 5 dias foi calculado pela diferença de valores entre comprimento radicular obtidos após 3 e 5 dias e comprimento radicular inicial, $\text{CL}_{3d} = C_{3d} - C_{0d}$ e $\text{CL}_{5d} = C_{5d} - C_{0d}$, respectivamente. Para cada tratamento foram avaliadas duas repetições de sete plantas, distribuídas de acordo com o delineamento inteiramente casualizado, em câmara de crescimento com temperatura diurna média de $27 \pm 3^\circ\text{C}$, noturna de $20 \pm 3^\circ\text{C}$ e fotoperíodo de 12 horas. Para cada tratamento com Al as médias de crescimento líquido radicular, $\text{CL}_{3d+\text{Al}}$ e $\text{CL}_{5d+\text{Al}}$, de cada uma das duas repetições foram divididas pela média global do crescimento líquido radicular sem Al, $\text{CL}_{3d-\text{Al}}$ e $\text{CL}_{5d-\text{Al}}$, sendo assim gerados dois valores percentuais de crescimento líquido relativo, CLR_{3d} e CLR_{5d} .

Notas para Injúria Radicular Visual (IRV) foram atribuídas após 5 dias de exposição ao Al de acordo com a escala de notas apresentada no capítulo anterior. Para visualizar a presença de indução do crescimento radicular, foram ajustados modelos de regressão linear com base nos valores de comprimento radicular obtido após 3 dias (C_{3d}) e 5 dias (C_{5d}) de exposição ao estresse de Al. A inclinação da reta obtida pela regressão linear foi adotada como índice de indução (índice b), onde $b=0$ indica ausência de indução do crescimento radicular e $b>0$ indica indução do crescimento radicular sob estresse de Al.

O primeiro componente principal (CP1), cuja estimativa foi descrita no capítulo 1, que explica 83% da variação de CLR_{3d} , CLR_{5d} , IRV e índice b foi adotado como índice geral de tolerância.

2.3. Sequenciamento e Genotipagem dos Polimorfismos na Região Alvo

O DNA genômico foi isolado de um bulk de tecido vegetal de cinco plantas por acesso, seguindo o protocolo baseado em CTAB (Saghai-Maroo et al. 1984). Trinta e oito

marcadores SSR do kit de referência de sorgo (http://sat.cirad.fr/sat/sorghum_SSR_kit) foram usados para caracterizar a estrutura de população. Os tamanhos dos alelos em pares de bases (pb) para os 209 acessos do CIRAD/ICRISAT foram fornecidos pelo Generation Challenge Program e sendo que os 45 acessos da Embrapa foram genotipados neste trabalho. Detalhes sobre condições das reações e correção para diferenças de tamanhos dos alelos entre laboratórios estão descritos no capítulo 1.

As regiões consideradas candidatas para conterem os polimorfismos causativos de efeitos fenotípicos entre SC283 e BR007 foram genotipadas ou seqüenciadas no painel de associação, de maneira a amostrar a região de 24,6kpb onde está localizado o loco *Alt_{SB}* (Figura 1). Informações sobre os primers estão descritas na Tabela 1. Um fragmento amplificado com o par de primers MJ3/MJ6 revelou um indel de 7pb e um SNP G→A na ORF6, que estão separados do gene *SbMATE* por um evento de recombinação e que por isso podem ser descartados como polimorfismos causativos. O fragmento amplificado com o par de primers JL51/JL52 revelou um SNP T→A identificado no exon 2 do gene *SbMATE* somente na linhagem SC566, enquanto que o par de primers JL55/JL56 amplificou seis SNPs e um indel localizados no intron 2 do gene *SbMATE*. A amplificação com o par de primers JL237/JL238 revelou o indel de 19pb localizado na ORF9 e o fragmento amplificado com o par de primers JL225/JL226 contém uma região da ORF10 que contém dois SNPs. Essa última região também está separada do gene *SbMATE* por um evento de recombinação sendo, portanto, descartada como polimorfismo causativo.

Tabela1. Primers adotados na amplificação de cada região amostrada.

Localização	Primers	Sequência dos primers 5'-3'
ORF6	MJ3	TCGGCTGTCGGCTCCCTAAACTCA
	MJ6	GCCGAGCGGAGGAGGAC
MITE	MJ23	CAGCGAATAATACTTTTAGTCATA
	MJ26	ACACCGCCTTTTCAGTTTTTACG
<i>SbMATE_E1</i>	JL51	GGCACGCACAGGCACAGTAACTTA
	JL52	AACAAGTGGCCAAGTGGGTGATCA
<i>SbMATE_I2</i>	JL55	GCCCGCGCTGCGCTACCTGA
	JL56	ATCCGAGGAAGCGCCGGAAT
ORF9	JL237	GCAGATTCTTTGCCAGGTGT
	JL238	CAGTAGCAGGCATGATGTCG
ORF10	JL225	TCTGGGCGGAATACAAGGTT
	JL226	TGGGAGTTCATGGTGGGTAAAA

MJ3/MJ6 Ciclo: 95°C-2 min/ 35x: 96°C-10 seg, 58°C-10 seg, 72°C-30 seg / 72°C-5 min
 MJ23/MJ26 Ciclo: 95°C-2 min / 30x: 94°C-1 min, 55°C-1min, 72°C-2 min / 72°C-10 min
 JL55/JL56 Ciclo: 95°C-2 min / 35x: 94°C-30 seg, 55°C-30 seg, 72°C-1 min / 72°C-5 min
 JL51/JL52 Ciclo: 95°C-2 min / 35x: 94°C-30 seg, 60°C-30 seg, 72°C- 90seg / 72°C-5 min
 JL237/JL238 Ciclo: 95°C-2 min / 35x: 94°C-30 seg, 55°C-30 seg, 72°C-1 min / 72°C-5 min
 JL225/JL226 Ciclo: 95°C-2 min / 35x: 94°C-30 seg, 55°C-30 seg, 72°C-1 min / 72°C-5 min

Para condução das reações de sequenciamento foram amplificados fragmentos de DNA em um volume final de 20µl contendo 30ng de DNA genômico, 20X tampão Taq, 2.0mM dNTP, 40mM MgCl₂, 10.0 pmol de cada primer, 5% de dimetilsulfóxido (DMSO) e 0.5U da enzima Taq polymerase (Phoneutria, Belo Horizonte). As amplificações foram conduzidas de acordo com os ciclos descritos na Tabela 1. A um total de 5µl de amostra contendo produtos de cada amplificação foram adicionadas 5U de Exonuclease I (ExoI) e 1U de Fosfatase alcalina de camarão (*Shrimp Alkaline Phosphatase*, SAP) e incubados a 37°C por 15 min seguidos por 80°C por 15 min. Os fragmentos amplificados previamente tratados com Exo/SAP foram usados como molde para o sequenciamento usando 2 µL de BigDye Terminator (Applied Biosystems, Foster City, CA), 2 µL de tampão 5X (Applied Biosystems, Foster City, CA) e 10 pmoles de cada primer em um volume final de 10 µL. As reações foram submetidas a 96°C por 20 seg, 50°C por 15 seg, 60°C por 4 min, repetidos por 30 vezes. O sequenciamento foi feito no sequenciador automático ABI3100 (Applied Biosystems, Foster City, CA). As sequências obtidas foram alinhadas usando os eletroferogramas utilizando-se o software SEQMAN (DNASTar, Madison, WI) e as sequências foram curadas manualmente. Os polimorfismos do MITE foram determinados via genotipagem com os primers MJ23 e MJ26 utilizando-se as condições de amplificação descritas na Tabela 1 e resolvendo-se os produtos de amplificação em gel de agarose 1% (m/v), em tampão TAE (40 mM Tris-acetato; 1 mM EDTA, pH 8,0) à 100V.

2.4. Desequilíbrio de Ligação

O desequilíbrio de ligação entre pares de locos polimórficos foi estimado pelo quadrado do coeficiente de correlação entre frequências alélicas (r^2 , Hill e Robertson, 1968) e coeficiente padronizado de desequilíbrio (D' , Lewontin, 1964) como implementado no software Tassel v.2.1 (<http://www.maizogenetics.net/bioinformatics/tasselindex.htm>). A significância foi calculada pelo teste exato de Fisher. Os cinco alelos revelados pelo MITE foram tratados como locos bialélicos para os cálculos de desequilíbrio de ligação e para as análises de associação. Valores de r^2 foram plotados em relação à distância (pb) entre locos e a queda do desequilíbrio de ligação foi avaliada usando modelos de deriva-recombinação e deriva-recombinação-mutação, como descritos em Remington et al (2001) utilizando-se um modelo de regressão não linear (PROC NLIN, SAS software). O primeiro modelo estima o valor esperado de r^2 considerando o equilíbrio deriva-recombinação, $E(r^2)=1/(1+C)$ onde $C = 4Nc$ (N é o tamanho efetivo da população e c a fração de recombinação entre locos) enquanto que no segundo modelo é adicionada a taxa de mutação $u=c$. O modelo não linear estima um coeficiente único, que corresponde ao parâmetro $4Nc$ por pb de distância entre locos.

2.5. Análise de Associação

2.5.1. Modelo Linear Misto

Os testes de associação entre tolerância ao Al e os polimorfismos candidatos foram conduzidos segundo quatro diferentes modelos (Yu et al. 2006). O primeiro modelo não inclui controle de estrutura de população tão pouco de parentesco (chamado de modelo *naive* ou simples). O modelo simples corresponde à $y = A\alpha + e$ onde, y é o vetor de observações fenotípicas; α é o vetor de efeitos fixos relacionado aos polimorfismos candidatos; A é uma matriz incidência de 0 e 1 relacionando y a α e e é o vetor de efeitos residuais. O segundo modelo incorpora o fator estrutura de população (Q) e pode ser definido como $y = Q\iota + A\alpha + e$ onde ι é o vetor de efeito fixo relacionado a estrutura de população e Q é a matriz de designação de membros nas populações (obtida pelo software STRUCTURE, Pritchard et al. 2000) relacionando y a ι . O terceiro controla parentesco, mas não estrutura de população, definido como $y = Zu + A\alpha + e$ onde u é o vetor de efeitos aleatórios relacionado com parentesco e Z é uma matriz incidência de 0 e 1 relacionando y a u . Finalmente, foi utilizado um modelo misto que controla simultaneamente estrutura de população e coancestralidade ($Q+K$), $y = Q\iota + A\alpha + Zu + e$, com cada um dos seus termos como definido acima. As variâncias dos efeitos aleatórios são expressas como $\text{Var}(u) =$

$2\mathbf{K}\mathbf{V}_g$, e $\text{Var}(\mathbf{e}) = \mathbf{R}\mathbf{V}_R$, onde $\mathbf{K}$ é a matriz 254 x 254 de coeficientes de parentesco baseada na proporção de alelos compartilhados que define o grau de covariância genética entre dois indivíduos, $\mathbf{R}$ é uma matrix 254 x 254 sendo os elementos fora da diagonal igual a zero e os elementos da diagonal igual ao número recíproco de observações fenotípicas para qual cada dado fenotípico foi obtido. $\mathbf{V}_g$ é a variância genética e $\mathbf{V}_R$ é a variância residual.

2.5.2. Estrutura de População e Parentesco

O software STRUCTURE versão 2.1 (Pritchard et al. 2000; <http://pritch.bsd.uchicago.edu/structure.html>) foi usado para estimar o número mais provável de subpopulações, sendo os quatros níveis de diferenciação ($k=4$, $k=6$, $k=8$ e $k=12$) identificados no capítulo 1 testados para o controle do Erro Tipo I devido ao fator estrutura de população na análise de associação. O programa gera uma matriz denominada Q que indica qual a proporção do genoma do indivíduo i é originado de cada uma das subpopulações. O número o mais provável de subpopulações foi avaliado considerando o modelo de miscigenação com frequência de alelos correlacionada, com o tamanho do burn-in de 100.000 e o tamanho da corrida de 1.000.000, como detalhado no capítulo 1.

Para a estimativa do coeficiente de parentesco, a matriz baseada na proporção de alelos compartilhados entre pares de acessos (D_{AC} , Jin e Chakraborty, 1993) foi obtida por $D_{AC} = 1 - \left(\frac{a}{2n}\right)$ onde a é o número de alelos em comuns entre os indivíduos comparados e n é o número de locos estudados. O uso dessa estimativa tem proporcionado controle mais eficiente de falsos positivos em estudos de associação (Weber et al. 2008) e foi superior a outros métodos testados nesse trabalho (dados não mostrados). A matriz foi obtida no software PowerMarker (Liu e Muse 2005).

2.5.3.Otimização do Modelo Linear Misto para Controle do Erro Tipo I

O desempenho do modelo simples e dos modelos que controlam diferentes níveis de relacionamento entre indivíduos, K (parentesco), Q (estrutura de população) e Q+K, foi testado para o controle de falso positivos (Erro Tipo I) com 38 marcadores *Simple Sequence Repeat* (SSR) não ligados e aleatoriamente distribuídos no genoma. Considerando o desequilíbrio de ligação esperado nessa população, que requeriria um número muito grande de marcadores para que se atingisse a saturação, somente 38 locos SSR resultam em uma probabilidade extremamente baixa, ou até mesmo desprezível, de associação com

características fenotípicas. Portanto, eles fornecem uma distribuição nula na qual os modelos descritos acima foram testados para controle do Erro Tipo I. Foram considerados alelos com frequência maior que 10% para as análises de performance dos modelos.

2.5.4. Influência das Diferentes Subpopulações e Classes de Tolerância ao Al nos Resultados da Análise Associativa

Devido às evidências de que a tolerância ao Alumínio em sorgo está relacionada com subpopulações formadas pelas raças caudatum e guinea (capítulo 1), foi estudada a influência de cada uma delas no resultado da análise de associação. Foram gerados três novos conjuntos de dados, retirando do conjunto original as subpopulações com predominância de ambas as raças. Em seguida, foram removidos do conjunto original, os acessos presentes nas subpopulações com predominância de acessos guinea e finalmente foram removidos do conjunto original os acessos das subpopulações com predominância de acessos caudatum. Foi também estudada a contribuição de cada uma das três classes de tolerância ao Al para detecção de associação. Para esse fim foram gerados outros três conjuntos de dados, removendo do conjunto de dados original os acessos classificados como sensíveis para a tolerância ao Al ($CLR_{5d} < 30\%$), em seguida os acessos classificados como intermediários ($30\% < CLR_{5d} < 80\%$) e finalmente, os acessos classificados como tolerantes ao Al ($CLR_{5d} > 80\%$). Para todos os conjuntos de acessos, o número mais provável de subpopulações foi avaliado, considerando o modelo de miscigenação com frequência de alelos correlacionada, com o tamanho do burn-in de 100.000 e o tamanho da corrida de 1.000.000. O número de k variou de 1 até 14 com cinco corridas independentes para cada k. O critério para detectar o mais provável valor de k foi o proposto por Evanno et al (2005) como descrito no capítulo 1.

2.6. Diversidade na Região Alvo

A diversidade nucleotídica foi obtida por
$$\pi = \sum_{i=1}^n \sum_{j=1}^i p_i p_j \pi_{ij}$$
 onde p_i e p_j são as frequências dos alelos (sequências) i e j , e π_{ij} é a proporção de nucleotídeos polimórficos quando a i -ésima e j -ésima sequências são comparadas. A taxa de mutação na população, foi obtida pelo estimador de Watterson (1975), $\theta_w = \frac{S}{a_n}$, onde S é o número de locos segregantes e $a_n = (1 + \frac{1}{2} + \dots + \frac{1}{n-1})$, onde n é o número de sequências avaliadas. O teste D de Tajima (Tajima 1989) foi usado para testar evidências de seleção na região do loco Alt_{SB} .

Os polimorfismos do MITE não foram considerados para essas análises devido à taxa de mutação diferencial com relação aos outros polimorfismos.

Os polimorfismos que apresentam associação para tolerância ao Al foram usados para identificação de haplótipos. Foi também calculada a probabilidade de ocorrência de eventos de recombinação intraloco usando o teste dos quatro gametas (Hudson e Kaplan, 1985). Todas as análises foram conduzidas no DnaSP (v.3.51; <http://www.ub.es/dnasp/>).

2.7. Análise Filogenética da Região Alvo

Foi contruída uma árvore de máxima parsimônia com os polimorfismos dos 30 haplótipos identificados na região de 24,6kpb. O loco referente do transposon MITE foi removido desta análise devido a sua natureza multialélica. Suporte para os clusters foram obtidos por 1000 bootstraps. As análises foram conduzidas no Software MEGA (Kumar et al. 2001).

3. Resultados

3.1. Padrão de Desequilíbrio de Ligação na Região do Gene Alt_{SB}

Foram identificados dezoito polimorfismos com frequência maior que 5% dentro das cinco regiões analisadas, compreendendo três polimorfismos de inserção/deleção (indels), cinco alelos do MITE e 10 *Single Nucleotide Polymorphisms* (SNPs).

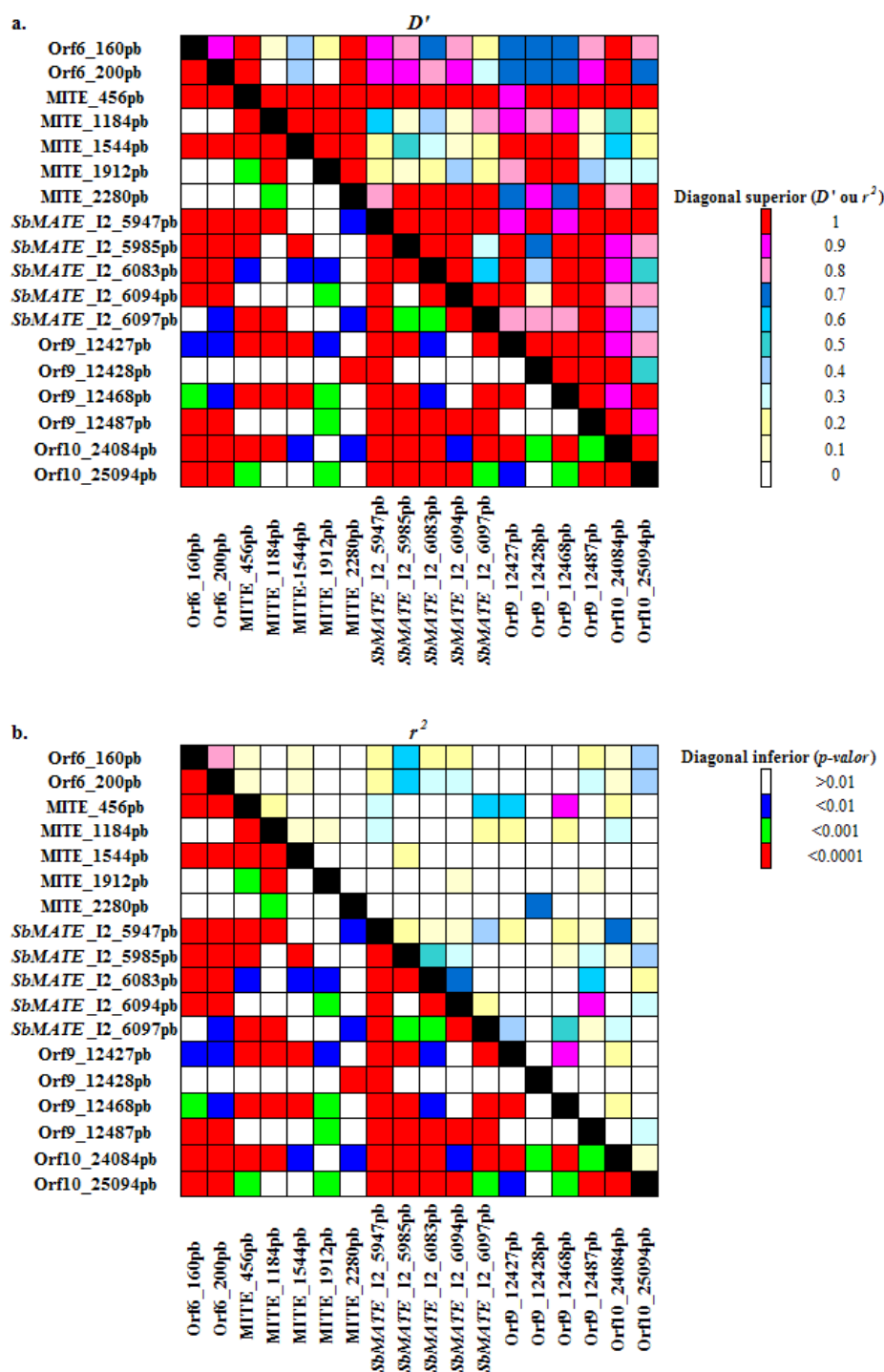


Figura 2. Matriz de desequilíbrio de ligação das cinco regiões analisadas. Foram considerados os locos com frequência maior que 5%. Cada quadrado acima da diagonal da matriz representa a estimativa de (a) D' e (b) r^2 para um dado par de locos. Na diagonal inferior estão apresentados os p-valor obtidos com o teste exato de Fisher. Os valores das estimativas de D' , r^2 e p-valor estão apresentados com código de cores no lado direito da Figura 2a.

As estimativas de desequilíbrio de ligação entre locos, obtidas com o coeficiente padronizado de desequilíbrio (D' , Figura 2a), foram muito superiores aos obtidos com r^2

(Figura 2b). Uma das possíveis razões para isso é o fato das estimativas D' serem normalizadas com base nas frequências alélicas, sendo, portanto, afetadas principalmente pelo histórico de recombinação na população (Flint-Garcia et al. 2003). Já as estimativas de r^2 , que são também afetadas pelas diferenças nas frequências alélicas entre locos, refletem, concomitantemente, o histórico de mutação e de recombinação na população. Assim, r^2 igual a um será alcançado somente quando as frequências alélicas dos locos comparados forem idênticas, o que não ocorreu frequentemente neste conjunto de dados (Figura 3).

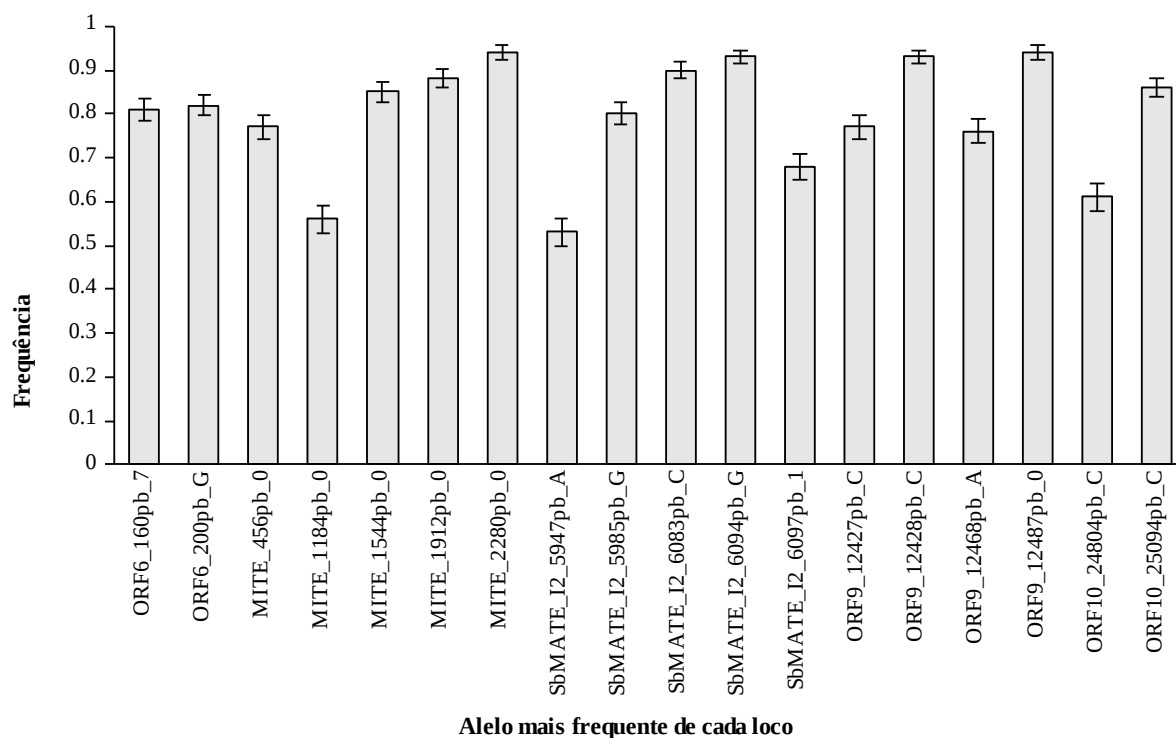


Figura 3. Distribuição de frequências do alelo mais frequente de cada loco analisado na região de 24,6kbp do loco *Alt_{SB}*. Desvio padrão (DP) das frequências foi assim calculado:

$$DP = \sqrt{\frac{p1(1 - p1)}{n}}, \text{ onde } p1 \text{ é a frequências dos alelo mais frequente de cada loco analisado e } n \text{ é o número de acessos avaliados.}$$

Com base nos valores de D' foram observados blocos de desequilíbrio de ligação na ORF6, no gene *SbMATE* e na ORF9 (Figura 2a). Entre os alelos do MITE, somente o alelo de 456pb apresentou desequilíbrio de ligação consistentemente elevado ($D' = 0.9$ até 1.0) com os demais polimorfismos. O alelo do MITE de 1544pb apresentou desequilíbrio de ligação completo com os alelos da ORF9 localizados a 12427, 12428 e 12468pb, enquanto que o alelo do MITE de 1912pb apresentou desequilíbrio de ligação completo com somente dois locos da ORF9, localizados a 12428 e 12468pb. Portanto, os padrões de desequilíbrio

obtidos com o MITE indicam pouca consistência com o fator distância física, o que possivelmente reflete a natureza particular desse loco com relação aos outros. A manutenção de blocos em desequilíbrio de ligação permitem a seleção de um número reduzidos de SNP (*tag* SNPs) para representar a região.

Visando estimar a extensão física do desequilíbrio de ligação na região do loco, o modelo deriva-recombinação foi ajustado aos dados (Figura 4a) e explicou uma maior fração da variação total do que o modelo que incorpora mutação (Figura 4b), indicando um declínio de $r^2 \cong 0.3$ ou menos para locos separados por 2500pb, e atingindo o equilíbrio ($r^2 \cong 0$) em distâncias superiores a 5000pb.

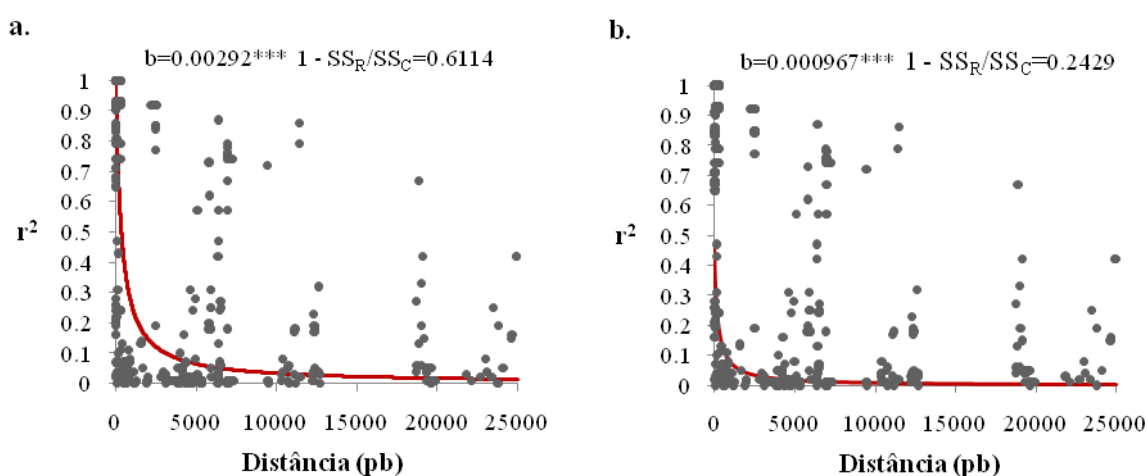


Figura 4. Decaimento de desequilíbrio de ligação (r^2) em função da distância física entre pares de locos. As curvas representam o ajuste de modelos de regressão nao-linear de r^2 em relação à distância física entre locos, com o modelo deriva-recombinação (a) e o modelo deriva-recombinação-mutação (b). Valores de R^2 são apresentados após a estimativa de b.

3.2. Estrutura de População e Controle do Erro Tipo I

A presença de sub-grupos com frequências alélicas diferentes dentro da população (Pritchard 2001) tem sido considerada uma fonte importante de falsos positivos nos estudos de associação, ou seja, marcadores não ligados encontrando-se em desequilíbrio de ligação com a característica em estudo. Assim, para reduzir o número de associações espúrias, deve-se identificar e controlar estatisticamente a estrutura de população.

Como esperado, os 254 acessos de sorgo não constituem uma única população em Equilíbrio de Hardy Weinberg (EHW), o que foi confirmado pelo desvio de todos os 38 locos SSR das proporções esperadas para a condição de equilíbrio (dados não mostrados), pelo padrão complexo de estruturação apresentado pelos acessos avaliados, como

demonstrado pelos quatro níveis de estruturação identificados pelo teste de Evanno et al (2005), correspondente a $k=4$, $k=6$, $k=8$ e $k=12$, discutidos no capítulo anterior. Finalmente, os valores do coeficiente de endogamia F_{ST} (Wright 1951) para cada par de subpopulações considerando $k=6$ (Tabela 3) também indicaram estruturação dentro do conjunto de 254 acessos de sorgo. Entre os quatro níveis de estruturação, $k=6$ foi escolhido pela eficiência de agrupamento e pelo sentido biológico dos seis grupos definidos (capítulo 1)

Tabela 3. Estimativas de F_{ST} baseadas nos 38 locos SSR considerando o número de subpopulações igual a seis ($k=6$), definido no capítulo 1

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	0	0.1818	0.2447	0.2605	0.1648	0.2225
Q2		0	0.1590	0.2283	0.1449	0.1898
Q3			0	0.1930	0.2249	0.2341
Q4				0	0.2375	0.1921
Q5					0	0.1982
Q6						0

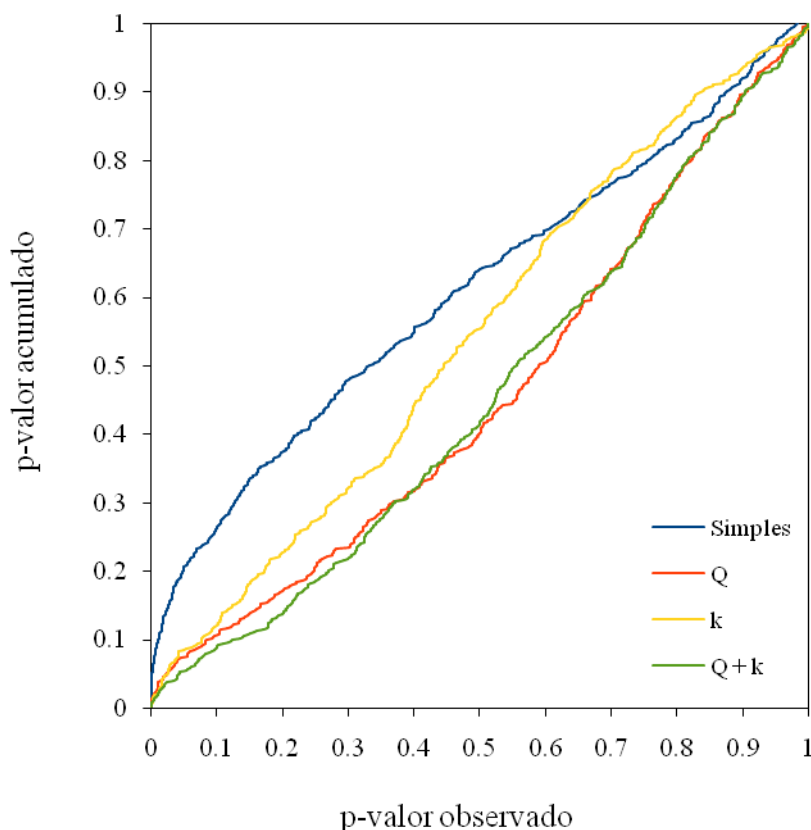
Uma vez detectada a presença de estrutura de população entre os 254 acessos de sorgo, procedeu-se à simulação do Erro Tipo I para os modelos Simples, Q, K e Q+K, considerando os quatro níveis de estruturação identificados pelo teste de Evanno et al (2005), correspondentes a $k=4$, $k=6$, $k=8$ e $k=12$ (Tabela 4).

Um maior número de falsos positivos foi observado para os modelos Simples, Q e K em todos os quatros níveis de estruturação testados, sendo que o modelo simples foi o que apresentou a pior performance (Tabela 4). O controle mais efetivo de falsos positivos foi alcançado com o modelo que inclui seis subpopulações e a matriz de parentesco obtida pelo cálculo da proporção dos alelos SSR compartilhados entre indivíduos. Dessa maneira, 5% dos dados apresentaram p-valor menor ou igual a 5% utilizando-se o modelo Q(6)+K em comparação com 7 e 8% de falsos positivos com os modelos Q e K, respectivamente, e 19% com o modelo simples.

Tabela 4. Porcentagem de associações significativas de acordo com a hipótese nula de independência entre os alelos SSR e as variáveis CLR_{3d}, CLR_{5d}, IRV, índice b e CP1 ao nível de significância de 5%.

	Q=4	Q=6	Q=8	Q=12
Simples	19	19	19	19
Q	11	7	7	11
K	8	8	8	8
Q + K	7	5	7	7

Os p-valores observados obtidos pela análise de associação entre os 38 locos SSR e as cinco variáveis CLR_{3d}, CLR_{5d}, IRV, índice b e CP1 foram então ordenados e plotados contra os p-valores acumulados para se estudar a distribuição de p-valores de cada modelo de acordo com hipótese nula (Figura 5). Foi possível observar uma distribuição não uniforme de p-valores para os modelos Simples, Q(6) e K, resultando em estimativas extremamente infladas de Erro Tipo I com relação ao modelo Q(6)+K. Isto indica que os modelos simples, Q e K não são os mais adequados para o mapeamento associativo nesse conjunto de acessos, uma vez que eles iriam resultar em um maior número de associações espúrias do que o modelo Q(6)+K, com Erro Tipo I bastante superior ao admitido. Entretanto o fator estrutura de população isoladamente apresenta maior contribuição que o fator parentesco no controle do Erro Tipo I neste conjunto de dados, como pode ser observado pela distribuição de p-valores apresentada na Figura 5. Assim, concluímos que o modelo linear misto, com seis subpopulações e matriz de parentesco baseada na proporção de alelos compartilhados controla adequadamente o número de associações falsas e pode seguramente ser adotado para relacionar os polimorfismos da região de 24,6kpb com tolerância ao Al.



Figura

5. Distribuições do Erro Tipo I obtidas com os modelos simples, Q (Q=6, considerando a estrutura de população com a presença de 6 sub-populações), K (baseado em parentesco entre indivíduos) e Q(6)+K (utilizando-se o modelo linear misto) para 38 locos SSR. Assumindo-se que os locos SSR distribuídos aleatoriamente não estão associados às características CLR_{3d}, CLR_{5d}, IRV, índice b e CP1 (Ho: ausência de efeito dos locos SSR), espera-se uma distribuição uniforme dos p-valores ao longo de uma linha diagonal no gráfico (p-valores observados próximos aos p-valores acumulados). Desvios desse padrão indicam baixa performance do modelo em reduzir o Erro Tipo I ou perda de poder estatístico para detectar associações.

3.3. Análise de Associação

As associações entre os 18 polimorfismos da região alvo (Figura 6) e as variáveis CLR_{3d}, CLR_{5d}, IRV, índice b e CP1 foram testadas de acordo com os modelos Simples, Q(6), K e Q(6)+K.

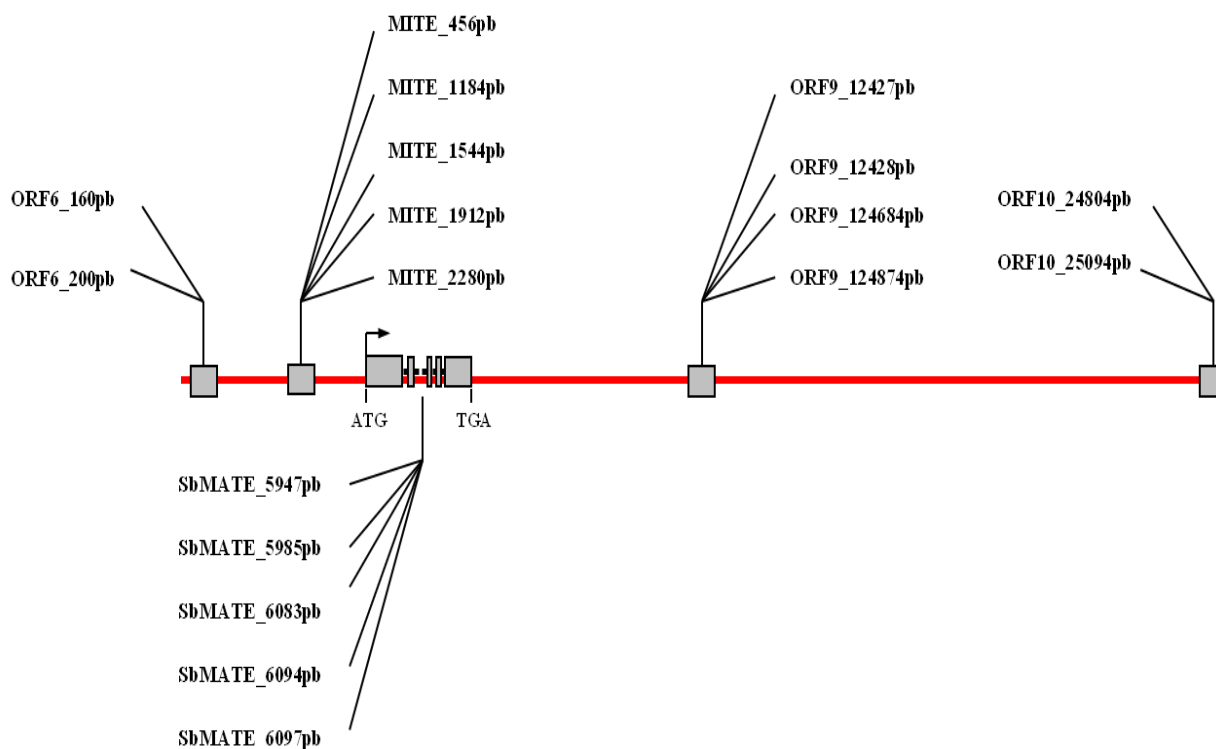


Figura 6. Distribuição dos 18 polimorfismo na região de 24,6kpb do loco *Alt_{SB}*.

Como esperado, os testes com os modelos Simples, Q(6) e K resultaram em um maior número de polimorfismos associados que o modelo Q(6)+K (Tabela 5). Foi também observado que, mesmo para locos que permaneceram associados com o modelo Q (6)+K, a porcentagem da variação fenotípica explicada foi bastante reduzida em relação aos outros modelos. Por exemplo, o loco *SbMATE_I2_6083pb* explicou 35.9% da variação do CP1 quando o modelo Simples foi considerado, 19.7% quando o modelo Q foi utilizado, 16.0% com o modelo K e 10.8% quando incorporou-se a correção para estrutura de população e parentesco por intermédio do modelo Q (6)+K. Sendo assim, a ausência de controle de estrutura de população e de parentesco entre indivíduos não só aumentou o número de associações espúrias mas provavelmente superestimou a porção da variação fenotípica explicada pelos locos associados. Por esta razão, doravante serão considerados somente os polimorfismos identificados de acordo com o modelo linear misto com seis subpopulações e matriz de coancestralidade obtida com o número de alelos compartilhados (Q(6)+K). A posição dos polimorfismos associados no contexto do loco *Alt_{SB}* estão mostrados na Figura 7 e estão localizados na ORF6, MITE, gene *SbMATE*, ORF9 e ORF10.

Tabela 5. Resultados dos testes de associação para polimorfismos distribuídos na região de 24,6kpb do loco *Alt_{SB}* (indicados na Figura 6) e fenótipos relacionados à tolerância ao Al. Para cada loco, estão apresentadas as classes de significância dos p-valores na primeira linha seguidos da porcentagem da variância fenotípica explicada para cada característica, com o alelo favorável indicado em negrito. Associações para tolerância ao Al foram conduzidas usando-se os modelos Simples, Q(6), K e Q(6)+K, com melhor performance do modelo Q(6)+K (sombreado claro). O loco *SbMATE_I2_6083pb* que explicou a maior porção da variância fenotípica para as várias características e que possivelmente está envolvido com a resposta indutiva (índice b) está marcado com sombreado escuro.

Locos	Polimorfismos	Simples					Q (6)					K					Q(6) + K				
		CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1
ORF6_160pb	0/7	*** 13.7	*** 14.7	*** 16.9	*** 14.7	*** 18.0	*** 4.5	*** 7.0	*** 8.1	*** 10.7	*** 9.2	*** 7.0	*** 8.0	*** 6.0	*** 6.0	*** 8.0	*** 3.0	*** 4.3	*** 4.5	*** 5.9	*** 5.4
ORF6_200pb	A/G	*** 13.5	*** 15.1	*** 16.5	*** 13.4	*** 17.7	*** 7.0	*** 9.8	*** 7.3	*** 7.8	*** 10.1	*** 8.0	*** 9.0	*** 6.0	*** 5.0	*** 8.0	*** 4.4	*** 6.1	*** 4.1	*** 4.4	*** 6.0
MITE_456pb	+/-	*** 5.6	*** 5.1	*** 3.7	*** 5.9	*** 6.1	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	*** 2.0	ns -	ns -	ns -	ns -	ns -	ns -
MITE_1184pb	+/-	ns -	ns -	*** 4.9	*** 5.1	ns -	ns -	*** 4.2	*** 5.9	*** 7.4	*** 5.8	*** 1.0	*** 3.0	*** 3.0	*** 4.0	*** 3.0	*** 1.7	*** 2.9	*** 3.1	*** 4.5	*** 3.7
MITE_1544pb	+/-	ns -	ns -	*** 5.2	*** 3.9	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
MITE_1912pb	+/-	*** 14.0	*** 18.6	*** 16.8	*** 22.8	*** 21.8	*** 5.6	*** 7.4	*** 9.8	*** 12.8	*** 10.8	*** 8.0	*** 11.0	*** 9.0	*** 13.0	*** 12.0	*** 3.4	*** 4.0	*** 4.9	*** 6.4	*** 5.8
MITE_2280pb	+/-	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
<i>SbMATE_I2_5947pb</i>	A/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
<i>SbMATE_I2_5985pb</i>	A/G	*** 11.9	*** 14.3	*** 20.3	*** 16.3	*** 18.9	ns -	*** 3.7	*** 4.0	*** 5.3	*** 4.5	*** 5.0	*** 7.0	*** 7.0	*** 6.0	*** 8.0	*** -	*** -	*** -	*** -	*** -
<i>SbMATE_I2_6083pb</i>	A/C	*** 22.3	*** 29.3	*** 33.8	*** 33.5	*** 35.9	*** 7.4	*** 14.1	*** 18.3	*** 16.5	*** 19.7	*** 11.0	*** 14.0	*** 14.0	*** 15.0	*** 16.0	*** 4.9	*** 8.3	*** 8.8	*** 13.5	*** 10.8
<i>SbMATE_I2_6094pb</i>	C/G	*** 18.4	*** 24.6	*** 20.4	*** 17.3	*** 24.4	*** 5.5	*** 9.8	*** 9.6	*** 10.6	*** 11.0	*** 9.0	*** 12.0	*** 8.0	*** 7.0	*** 11.0	*** 4.2	*** 6.3	*** 4.1	*** 4.8	*** 6.1
<i>SbMATE_I2_6097pb</i>	0/1	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12427pb	C/T	*** 3.3	ns -	*** 4.0	*** 4.9	*** 4.6	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns 2.0	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12428pb	C/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12468pb	A/G	*** 3.6	*** 3.5	*** 3.7	*** 5.0	*** 4.7	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12487pb	19/0	*** 24.5	*** 18.5	*** 17.4	*** 16.6	*** 23.3	*** 6.2	*** 11.2	*** 10.3	*** 12.7	*** 12.5	*** 1.0	*** 13.0	*** 7.0	*** 8.0	*** 11.0	*** 5.1	*** 7.7	*** 4.7	*** 6.5	*** 7.5
ORF10_24804pb	C/T	ns -	ns -	ns -	ns -	ns -	*** 3.5	*** 4.6	ns -	ns -	*** 3.3	ns -	ns -	ns -	ns -	ns -	ns 2.6	ns 3.4	ns -	ns -	ns -
ORF10_25094pb	T/C	*** 5.8	*** 7.0	*** 9.8	*** 6.5	*** 8.8	ns -	ns -	ns -	ns -	ns -	*** 3.0	*** 3.0	*** 3.0	*** 2.0	*** 3.0	ns -	ns -	ns -	ns -	ns -

ns=não significativo, *** = $P < 0.001$. A porcentagem da variação fenotípica explicada, está apresentada para cada polimorfismo significativo.

CLR_{3d}: crescimento líquido relativo após 3 dias de exposição ao Al.

CLR_{5d}: crescimento líquido relativo após 5 dias de exposição ao Al.

IRV: injúria radicular visual.

Índice b: índice de indução do crescimento radicular obtido entre 3 e 5 dias de exposição ao Al.

CP1: primeiro componente da análise de componentes principais que responde por 83% da variação das variáveis CLR_{3d} , CLR_{5d} , IRV e índice b.

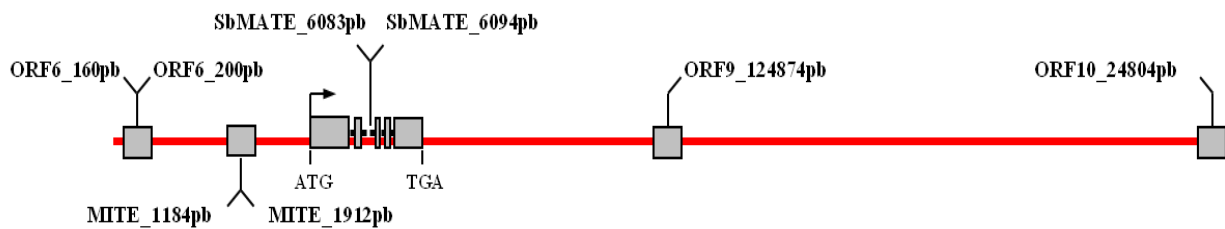


Figura 7. Localização dos polimorfismos associados na região de 24,6kpb do loco *Alt_{SB}*.

Os locos ORF6_160pb e ORF6_200pb, que estão em desequilíbrio de ligação de 0.9 (Figura 8), apresentaram associação significativa para todas as variáveis. Dentre os alelos do MITE, aqueles de 1184 e 1912pb apresentaram associações significativas e o desequilíbrio de ligação entre esses alelos e os demais polimorfismos não foi significativo. Os locos do segundo intron do gene *SbMATE*, *SbMATE_I2_6083pb* e *SbMATE_I2_6094pb*, apresentaram desequilíbrio de ligação completo entre si, provavelmente devido à proximidade física (esses dois locos estão separados por apenas 11pb), resultando em associações altamente significativas para ambos. Entretanto, o SNP A→C (*SbMATE_I2_6083pb*) explicou a maior parte da variação fenotípica para as variáveis estudadas, sugerindo a possibilidade de efeito funcional distinto para esse loco. Uma resposta diferenciada foi observada para o índice de indução (índice b), variável para a qual esse polimorfismo atingiu o mais alto valor, 13.5% da variação fenotípica explicada.

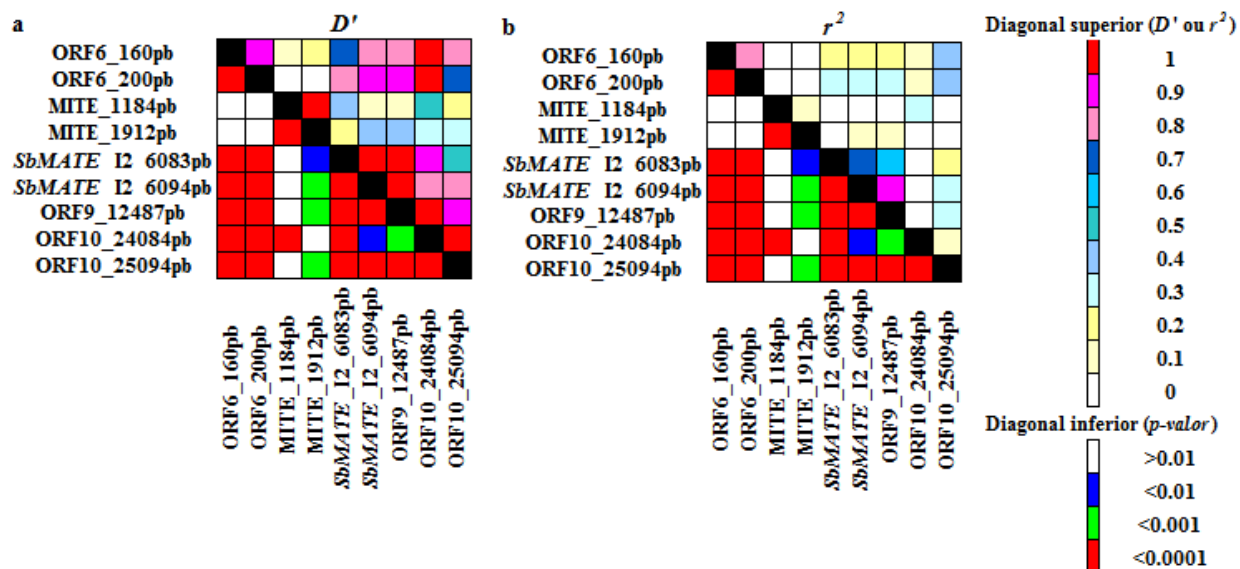


Figura 8. Matriz de desequilíbrio de ligação entre os polimorfismos associados na região de 24,6kpb do loco *Alt_{5B}*. Foram considerados os locos com frequência maior que 5%. Cada quadrado na matriz de desequilíbrio de ligação representa a comparação de cada par de locos , com o (a) D' e o (b) r^2 apresentados na diagonal superior e a significância do teste exato de Fisher apresentado na diagonal inferior. Os códigos de cores para D' , r^2 e p -valor estão apresentados no lado direito.

Dentre os polimorfismos da ORF9, somente o indel de 19pb (ORF9_12487pb) está em desequilíbrio de ligação completo com os SNPs associados do *SbMATE* e com o loco ORF10_24804pb. O SNP C→T da ORF10 (ORF10_24804pb) apresentou associação significativa somente para as variáveis CLR_{3d} e CLR_{5d} e em desequilíbrio de ligação completo com o indel de 19pb (ORF9_12487pb) e com os locos da ORF6, ORF6_160pb ORF6_200pb.

O poder estatístico para detecção de associações depende fundamentalmente da herdabilidade da característica, do tamanho da amostra, do balanço entre o efeito do QTN e a sua frequência alélica bem como da extensão do desequilíbrio de ligação entre outros SNPs e o QTN (Silva e Ball, 2007). No capítulo 1 foi mostrado que a tolerância ao Al é um evento raro em sorgo, com 6% dos acessos apresentando tolerância alta, 14% com tolerância intermediária e 80% apresentando sensibilidade extrema ao Al. No entanto, associações altamente significativas foram detectadas para polimorfismos na região do gene *Alt_{5B}*. Esses resultados levantam uma questão interessante: qual seria o pool gênico nesse painel que mais contribui para as associações detectadas? Para responder essa questão, os testes de associação foram refeitos, dessa vez eliminando, de forma sequencial, grupos de acessos definidos com base em

dois critérios, estrutura de população e tolerância ao Al. As associações previamente detectadas com o conjunto completo dos dados foram significativamente alteradas quando foram removidos do conjunto de acessos original, aqueles presentes nas subpopulações com predominância de ambas as raças, caudatum e guinea. Com exceção das inserções de 1184pb e 2280pb do MITE, os demais polimorfismos não apresentaram associações significativas. Esse resultado pode ser atribuído à importância de ambas as raças para a característica, mas também pode ser consequência da redução do tamanho da população ($n=63$) e perda de poder estatístico para detecção de associações (Tabela 6). Entretanto, com base nos resultados do capítulo 1, que mostraram a importância das raças guinea e caudatum para a tolerância ao Al, a primeira hipótese pode ser considerada razoável.

Em geral, houve um aumento notável na porcentagem da variação fenotípica explicada para cada polimorfismo associado quando as subpopulações com predominância da raça caudatum foram retiradas da análise, mantendo-se as demais. Entretanto, a variação explicada pelos polimorfismos associados apresentou padrão mais próximo ao da análise realizada no conjunto completo quando os acessos presentes nas subpopulações com predominância da raça guinea foram retirados, sendo mantidos os demais. Adicionalmente, a significância obtida com o 25904pb na ORF10 e a inserção de 1912 pb do MITE sugerem um efeito residual de estrutura de população causando as associações.

A contribuição superior das populações das raças guinea revelada pela análise de associação pode também refletir o fato dos polimorfismos amostrados terem sido previamente identificados na linhagem tolerante da raça guinea, SC283. Dessa maneira, polimorfismos importantes entre membros da raça caudatum podem ter sido omitidos.

Tabela 6. Resultados dos testes de associação para locos distribuídos na região de 24,6kpb do loco *Alt_{SB}* obtidos para o painel completo de acessos, painel completo exceto ambas as subpopulações Guinea e Caudatum e painel completo exceto cada uma dessas subpopulações separadamente. As subpopulações foram retiradas com base na análise de estrutura de população apresentada no Capítulo 1. Para cada loco, estão apresentados os p-valores na primeira linha seguidos da porcentagem da variância fenotípica explicada para cada característica, com o alelo favorável indicado em negrito. Associações para tolerância ao Al foram conduzidas usando o modelo linear misto e com o painel reduzido, redefinindo-se a matriz Q para cada caso para melhor controle do Erro Tipo I: com o conjunto de acessos completo, retirando as subpopulações com predominância das raças Caudatum e Guinea (Q=3), retirando as subpopulações com predominância da raça Caudatum (Q=9) e retirando as subpopulações com predominância da raça Guinea (Q=5). Associações significativas para cada conjunto de dados estão sombreadas.

Locos	Polimorfismos	Conjunto completo (n =244)					Acessos caudatum e guinea removidos (n = 63)					Acessos caudatum removidos (n =146)					Acessos guinea removidos (n =180)				
		CLR _{3d}	CLR _{5d}	IRV	índice b	CPI	CLR _{3d}	CLR _{5d}	IRV	índice b	CPI	CLR _{3d}	CLR _{5d}	IRV	índice b	CPI	CLR _{3d}	CLR _{5d}	IRV	índice b	CPI
ORF6_160pb	0/7	*** 3.0	*** 4.3	*** 4.5	*** 5.9	*** 5.4						*** 9.9	*** 3.4	*** 5.2	*** 7.3	*** 8.2	*** 5.2	*** 4.2	*** 4.0	*** 5.5	*** 4.3
ORF6_200pb	A/G	*** 4.4	*** 6.1	*** 4.1	*** 4.4	*** 6.0						*** 13.2	*** 4.2	*** 4.4	*** 8.7	*** 11.8	*** 6.4	*** 2.9	*** 3.1	*** 5.3	*** 4.7
MITE_456pb	+/-	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	*** 3.7	ns -	ns -	ns -	ns -	ns -	ns -	ns -
MITE_1184pb	+/-	*** 1.7	*** 2.9	*** 3.1	*** 4.5	*** 3.7	ns -	ns -	* 4.0	*** 8.0	* 4.0	ns -	ns -	*** 4.1	*** 3.8	ns -	ns -	ns -	*** 4.5	ns -	ns -
MITE_1544pb	+/-	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
MITE_1912pb	+/-	*** 3.4	*** 4.0	*** 4.9	*** 6.4	*** 5.8	ns -	ns -	ns -	ns -	ns -	*** 19.9	*** 14.0	*** 23.1	*** 21.7	*** 17.4	ns -	ns -	ns -	ns -	ns -
MITE_2280pb	+/-	ns -	ns -	ns -	ns -	ns -	ns -	*** 7.0	*** 8.0	*** 17.0	*** 12.0	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
SbMATE_I2_5947pb	A/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
SbMATE_I2_5985pb	A/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	*** 10.8	*** 5.1	*** 5.1	*** 8.1	*** 8.5	ns -	ns -	*** 3.9	*** 3.7	ns -
SbMATE_I2_6083pb	A/C	*** 4.9	*** 8.3	*** 8.8	*** 13.5	*** 10.8	ns -	ns -	ns -	ns -	ns -	*** 18.4	*** 10.6	*** 8.9	*** 14.8	*** 15.7	*** 9.7	*** 11.1	*** 16.3	*** 13.6	*** 6.7
SbMATE_I2_6094pb	C/G	*** 4.2	*** 6.3	*** 4.1	*** 4.8	*** 6.1	ns -	ns -	ns -	ns -	ns -	*** 22.3	*** 11.3	*** 9.7	*** 17.1	*** 19.4	*** 6.1	ns -	*** 4.2	*** 5.7	*** 4.7
SbMATE_I2_6097pb	0/1	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12427pb	C/T	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12428pb	C/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12468pb	A/G	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -
ORF9_12487pb	19/0	*** 5.1	*** 7.7	*** 4.7	*** 6.5	*** 7.5						*** 20.5	*** 7.0	*** 7.2	*** 13.8	*** 18.3	*** 8.0	*** 3.9	*** 6.4	*** 7.6	*** 6.1
ORF10_24804pb	C/T	*** 2.6	*** 3.4	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	ns -	*** 6.5	*** 3.7	ns -	*** 5.6	*** 4.5
ORF10_25094pb	T/C	ns -	ns -	ns -	ns -	ns -						*** 11.8	*** 4.2	*** 5.7	*** 8.4	*** 10.3	ns -	ns -	ns -	ns -	ns -

ns=não significativo, *** = $P < 0.001$. A porcentagem da variação fenotípica explicada está apresentada para cada polimorfismo significativo.

Dentre os 254 acessos de sorgo estudados, 80% apresentaram $CLR_{5d} < 30\%$ e foram considerados sensíveis, 14% apresentaram CLR_{5d} entre 30 e 80% e foram considerados de tolerância intermediária e somente 6% apresentaram $CLR_{5d} > 80\%$ e foram considerados tolerantes. Mesmo com a baixa frequência de acesso tolerantes, a detecção de associações significativas foi possível, o que nos levou a estudar a influência de cada classe de tolerância nos resultados de associação. Quando a análise foi conduzida na ausência dos acessos classificados com sensíveis houve perda de significância para a maioria dos locos previamente associados. Esse resultado não foi surpreendente, uma vez que a variabilidade fenotípica para a tolerância ao AI sofre redução acentuada com a retirada da classe sensível (Figura 9). Novamente, não pode ser descartado um possível efeito de tamanho populacional ($n=53$) causando perda no poder de detecção das associações.

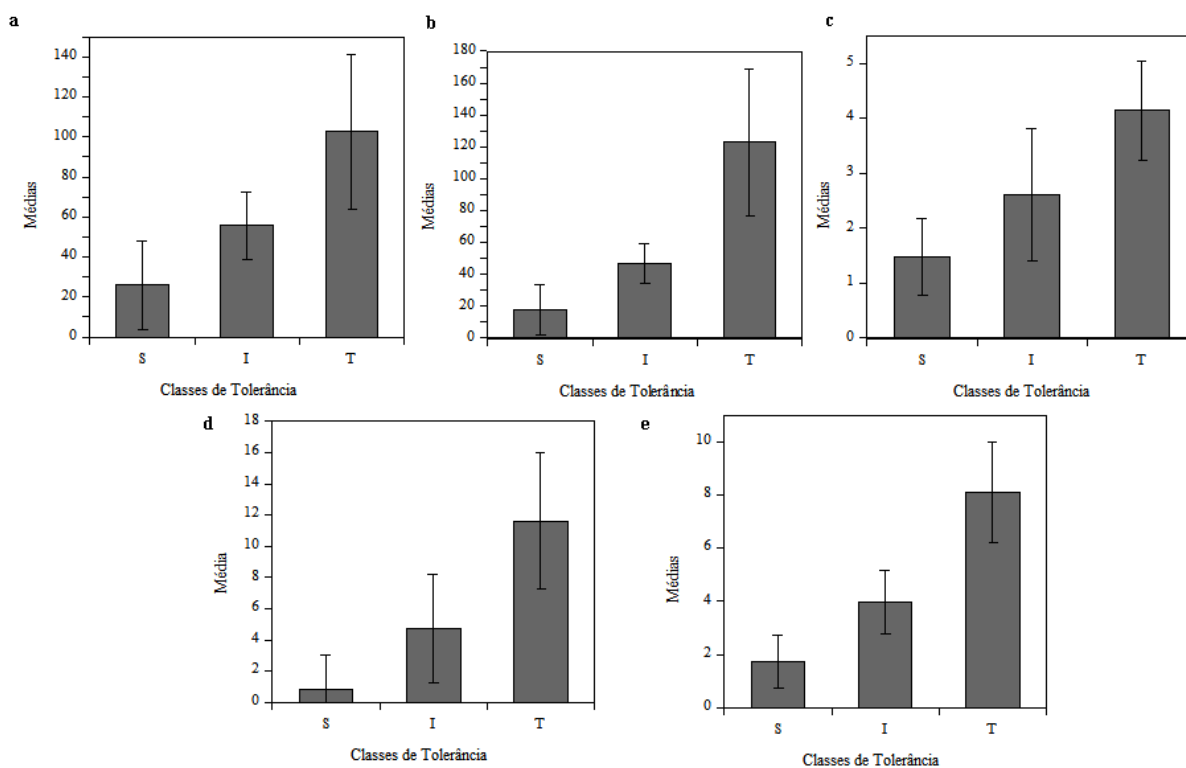


Figura 9. Valores médio das classes sensíveis (S), tolerantes intermediários (I) e tolerantes (T) para cada uma das cinco variáveis estudadas: CLR_{3d} (a), CLR_{5d} (b), IRV(c), índice b (d) e CP1 (e). Barras indicam desvio padrão.

Quando os acessos de tolerância intermediária foram removidos, associações foram detectadas, pois foram mantidas classes altamente contrastantes para as várias variáveis. Foi

impressionante o aumento da variação fenotípica explicada pelos polimorfismos quando os acessos de tolerância intermediária foram removidos da análise. Finalmente, quando os acessos tolerantes foram removidos da análise, associações foram verificadas, mas a variação explicada por cada polimorfismo foi dramaticamente reduzida quando comparada ao painel resultante da exclusão dos acessos intermediários.

Tabela 7. Resultados dos testes de associação para locos distribuídos na região de 24,6kbp do loco *Alt_{SB}* obtidos para o painel completo de acessos, painel completo exceto cada uma dos acessos sensíveis ($CLR_{5d} < 30\%$), acessos intermediários ($30\% < CLR_{5d} < 80\%$) e acessos tolerantes ($CLR_{5d} > 80\%$) separadamente. Para cada loco, estão apresentados os p-valores na primeira linha seguidos da porcentagem da variância fenotípica explicada para cada característica, com o alelo favorável indicado em negrito. Associações para tolerância ao Al foram conduzidas usando o modelo linear misto e com o painel reduzido, redefinindo-se a matriz Q para cada caso para melhor controle do Erro Tipo I, com o conjunto de acessos completo (Q=6), retirando os acessos sensíveis (Q=2), retirando os acessos intermediários (Q=7) e retirando os acessos tolerantes (Q=6). O resultado para o loco *SbMATE_6083pb* está sombreado

Locos	Polimorfismos	Conjunto completo (n=244)					Acessos sensíveis removidos (n=53)					Acessos intermediários removidos (n=209)					Acessos tolerantes removidos (n=227)				
		CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1	CLR _{3d}	CLR _{5d}	IRV	índice b	CP1
ORF6_160pb	0/7	***	***	***	***	***	ns	ns	ns	ns	ns	***	***	***	***	***	ns	***	ns	***	***
ORF6_200pb	A/G	3.0	4.3	4.5	5.9	5.4	-	-	-	-	-	6.0	7.0	3.0	4.0	6.0	-	3.0	-	5.0	3.0
MITE_456pb	+/-	***	***	***	***	***	ns	***	ns	ns	ns	***	***	***	***	***	ns	***	ns	ns	***
MITE_1184pb	+/-	4.4	6.1	4.1	4.4	6.0	-	9.0	-	-	-	9.0	10.0	5.0	5.0	8.0	-	3.0	-	-	2.0
MITE_1544pb	+/-	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	***	ns
MITE_1912pb	+/-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.0	-
MITE_2280pb	+/-	***	***	***	***	***	ns	ns	ns	***	ns	***	***	***	***	***	ns	***	***	***	***
<i>SbMATE_I2_5947pb</i>	A/G	1.7	2.9	3.1	4.5	3.7	-	-	-	11.0	-	3.0	4.0	4.0	5.0	5.0	-	3.0	2.0	3.0	3.0
<i>SbMATE_I2_5985pb</i>	A/G	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
<i>SbMATE_I2_6083pb</i>	A/C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>SbMATE_I2_6094pb</i>	C/G	***	***	***	***	***	ns	ns	ns	***	ns	***	***	***	***	***	***	***	***	***	***
<i>SbMATE_I2_6097pb</i>	0/1	4.9	8.3	8.8	13.5	10.8	-	-	8.0	-	10.0	12.0	15.0	18.0	16.0	18.0	2.0	6.0	3.0	8.0	6.0
ORF9_12427pb	C/T	***	***	***	***	***	ns	***	ns	ns	ns	***	***	***	***	***	***	***	ns	***	***
ORF9_12428pb	C/G	4.2	6.3	4.1	4.8	6.1	-	9.0	-	-	-	11.0	13.0	11.0	7.0	12.0	2.0	5.0	-	4.0	3.0
ORF9_12468pb	A/G	ns	ns	ns	ns	ns	ns	***	ns	ns	ns	***	***	ns	ns	ns	ns	ns	ns	ns	ns
ORF9_12487pb	19/0	-	-	-	-	-	-	9.0	-	-	-	2.0	2.0	-	-	-	-	-	-	-	-
ORF10_24804pb	C/T	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
ORF10_25094pb	T/C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

ns=não significativo, *** = $P < 0.001$. A porcentagem da variação fenotípica explicada está apresentada para cada polimorfismo significativo.

3.4. Diversidade na Região Alvo

Um total de 1667 pares de bases foram sequenciados para as análises de associação, que resultaram na identificação de 75 locos polimórficos, equivalendo a um polimorfismo a cada 22pb (Tabela 8). Fora da região codante (primeiro exon, *SbMATE_E1*) e do segundo intron do gene *SbMATE* (*SbMATE_I2*) os polimorfismos foram menos frequentes, com 1 polimorfismo a cada 66pb. Entretanto, a diversidade nucleotídica (δ) no gene *SbMATE* variou de 1.61/kbp a 3.8/kbp enquanto que fora do gene os valores de δ foram de 8.03/kbp. A taxa de mutação populacional, calculada por $\hat{\pi}_w$ foi de 2.03/kbp na região completa, sendo mais alta nas demais regiões, variando de 4.19/kbp a 9.92/kbp. Como os valores de δ foram sempre menores que os $\hat{\pi}_w$, o teste de D de Tajima foi negativo para todas as regiões. Entretanto somente para o exon 1 do gene *Alt_{SB}* foi observada significância para D, ($D = -1.77455$, $p \leq 0.05$).

Tabela 8. Diversidade nucleotídica entre os acessos de sorgo calculada para a região completa, segundo intron do *SbMATE*, segundo exon do *SbMATE* e as demais ORF6+ORF9+ORF10.

Região Genômica	N	Tamanho (pb)	S	δ (10^{-3})	$\hat{\pi}_w$ (10^{-3})	D Tajima
Completa	161	1667	75	1.10	2.03	-1.42595
<i>SbMATE_I2</i>	216	387	22	3.80	6.72	-1.17294
<i>SbMATE_E1</i>	219	357	39	1.61	4.19	-1.77455*
ORF6+ORF9+ORF10	164	923	14	8.03	9.92	-0.46467

N = número de acessos; S = número de sítios segregantes; δ = diversidade nucleotídica e, $\hat{\pi}_w$ = estimador de Waterson (1975) de $4N_e\mu/bp$.

Os 10 polimorfismos associados foram organizados em trinta haplótipos, dos quais treze apresentaram baixa frequência (0.5%), tendo sido representados por somente um acesso (Tabela 9). Oito haplótipos (*Alt_{SB}_2*, *Alt_{SB}_8*, *Alt_{SB}_12*, *Alt_{SB}_19*, *Alt_{SB}_20*, *Alt_{SB}_21*, *Alt_{SB}_22*, *Alt_{SB}_23* e *Alt_{SB}_27*) estão presentes em acessos tolerantes, $CLR_{5d} > 80\%$. Foi observada uma alta frequência de acessos sensíveis para os haplótipos *Alt_{SB}_1* (0.98) e *Alt_{SB}_2* (0.89) e, consequentemente, baixa frequência de acessos intermediários ou tolerantes. Portanto, é possível que a tolerância apresentada por estes acessos não esteja envolvida com o loco *Alt_{SB}*, mas com outros genes de tolerância. Marcadores flanqueando o loco *Alt_{SB}* não apresentaram ligação significativa para tolerância ao Al na população de retrocruzamento derivada da linhagens BR007 x IS29691. As linhagens IS29691 (tolerante) e BR007 (sensível) são ambas portadores do haplótipo *Alt_{SB}_2*, reforçando a hipótese de que a tolerância apresenta por linhagens portadoras destes haplótipos não são condicionadas pelo loco *Alt_{SB}*.

Um evento de recombinação entre as posições 12487 e 24801pb identificado no haplótipo *Alt_{SB}_21* pode ter sido responsável pela adição do SNP C na posição 24804pb da ORF10. Note-se que a linhagem altamente tolerante, IS23452, com CLR_{5d} de 163% e a linhagem BR012 com CLR_{5d} de 16% são portadoras desse haplótipo (Anexo1). Marcadores flanqueando o loco *Alt_{SB}* não apresentaram ligação significativa para tolerância ao Al na população de retrocruzamento derivada da linhagem IS23142 x BR007 (capítulo 1), indicando que a tolerância apresentada pela linhagem IS23142 pode ser devida a outros locos de tolerância. Ainda que a linhagem BR012 seja sensível ao Al, Caniato et al (2007) mostraram que o alelo *Alt_{SB}* presente em BR012 tem efeito fenotípico mais pronunciado do que aquele do padrão de sensibilidade, BR007, o que permitiu a detecção de ligação para marcadores flanqueando o loco *Alt_{SB}* em uma população F₂ derivada do cruzamento BR012 x BR007. A tolerância ao Al apresentada pela linhagem BR012, ainda que residual, pode ter sido adquirida pela adição do alelo C na posição 24804pb por meio de um evento de recombinação putativo tendo como um dos parentais um haplótipo tolerante.

Um dos haplótipos raros, *Alt_{SB}_12*, apresentou o mesmo evento de recombinação putativo que o *Alt_{SB}_21* entre as posições 5985 e 24804pb. Entretanto não podemos afirmar que este evento de recombinação é unicamente responsável pela tolerância da linhagem IS25077, única representante desse haplótipo (Anexo1) e que apresenta CLR_{5d} de 120%, uma vez que este haplótipo apresenta outros alelos favoráveis na região estudada.

As linhagens 3DX, 5DX, 9DX, SC175 e SC112 são todas portadoras do haplótipo *Alt_{SB}_20* (Anexo1). Caniato et al (2007) mostraram que a tolerância apresentada por 3DX, 9DX e SC175 é condicionada em grande parte pelo loco *Alt_{SB}* enquanto que o mesmo não ocorre para 5DX e SC112 (Caniato et al. 2007). Uma possível explicação é que o(s) polimorfismo(s) funcionais que diferenciam 3DX/9DX/SC175 de 5DX/SC112 podem estar em frequência muito baixa para serem detectados por associação ou podem não ter sido amostrados nesse estudo.

Tabela 9. Haplótipos identificados na região de 24,6kpb do loco *Alt_{SB}*. As posições dos intervalos de recombinação foram estimadas pelo teste dos quatro gametas (Hudson e Kaplan, 1985). A atribuição dos eventos de recombinação putativos aos haplótipos foi feita assumindo-se que o haplótipo ancestral corresponde a *Alt_{SB}_2* (portanto, todos os haplótipos recombinantes foram definidos com base no haplótipo *Alt_{SB}_2*). As barras finas identificam a recombinação entre os alelos dos locos ORF6_160pb e ORF6_200pb. As barras grossas indicam os intervalos de recombinação entre os alelos dos locos ORF6_200pb e *SbMATE*_5947pb e entre os alelos dos locos *SbMATE*_5985pb e ORF10_24804pb. À direita estão apresentadas as distribuição dos haplótipos em relação às classes de tolerância, classificação racial e subpopulações.

Haplótipo	ORF6_160pb ORF6_200pb MITE <i>SbMATE</i>_I2_5947pb <i>SbMATE</i>_I2_5985pb <i>SbMATE</i>_I2_6083pb <i>SbMATE</i>_I2_6094pb OR9_12487pb ORF10_24804pb ORF10_25094pb										Frequência	Tolerância ^a			Raça ^b						Subpopulação ^c					
	0/7	A/G		A/G	A/G	A/C	C/G	19/0	C/T	T/C		S	I	T	G	C	K	B	D	Gm	Q1	Q2	Q3	Q4	Q5	Q6
<i>Alt_{SB}_1</i>	7	G	456	A	G	C	G	0	C	C	23.2	48	1	-	2	18	-	8	8	-	-	20	3	-	25	2
<i>Alt_{SB}_2</i>	7	G	1184	G	G	C	G	0	T	C	29.4	55	6	1	9	7	14	5	1	-	-	7	2	34	3	14
<i>Alt_{SB}_3</i>	7	G	1184	G	G	C	G	0	C	C	5.2	10	1	-	3	2	1	3	-	-	4	3	-	1	3	-
<i>Alt_{SB}_4</i>	0	A	1184	A	A	C	G	0	C	C	1.9	4	-	-	2	-	-	-	-	-	3	-	-	-	1	-
<i>Alt_{SB}_5</i>	0	A	1184	A	A	C	G	0	C	T	1.9	4	-	-	4	-	-	-	-	-	4	-	-	-	-	-
<i>Alt_{SB}_6</i>	0	G	1184	A	A	C	G	0	C	C	0.5	1	-	-	-	1	-	-	-	-	-	2	-	-	-	-
<i>Alt_{SB}_7</i>	7	G	1184	A	G	C	G	0	C	C	0.5	1	-	-	-	-	1	-	-	-	-	-	-	-	1	-
<i>Alt_{SB}_8</i>	0	A	1184	A	A	A	C	19	C	C	0.5	-	-	1	1	-	-	-	-	-	-	-	-	-	-	1
<i>Alt_{SB}_9</i>	7	G	1184	A	A	C	G	0	T	C	0.5	1	-	-	-	-	-	-	1	-	-	-	-	1	-	-
<i>Alt_{SB}_10</i>	0	G	1184	A	G	C	G	0	C	T	0.5	-	1	-	-	-	-	-	-	-	-	-	-	-	1	-
<i>Alt_{SB}_11</i>	0	A	1184	A	A	A	C	19	C	T	1.4	1	2	-	-	-	-	-	-	3	3	-	-	-	-	-
<i>Alt_{SB}_12</i>	0	A	1184	A	G	C	G	0	C	C	0.5	-	-	1	1	-	-	-	-	-	1	-	-	-	-	-
<i>Alt_{SB}_13</i>	7	A	1184	A	A	C	G	0	C	T	0.5	1	-	-	-	-	-	-	-	-	1	-	-	-	-	-
<i>Alt_{SB}_14</i>	0	A	1184	G	G	C	G	0	C	C	0.5	-	1	-	1	-	-	-	-	-	1	-	-	-	-	-
<i>Alt_{SB}_15</i>	7	A	1184	A	A	A	C	19	C	T	0.5	1	-	-	-	-	-	-	-	-	1	-	-	-	-	-
<i>Alt_{SB}_16</i>	0	A	1544	A	A	C	G	0	C	C	1.4	2	1	-	2	-	-	-	-	1	3	-	-	-	-	-
<i>Alt_{SB}_17</i>	7	G	1544	G	G	C	G	0	T	C	2.4	3	2	-	1	1	1	-	-	1	-	1	1	1	-	2
<i>Alt_{SB}_18</i>	0	A	1544	A	A	C	G	0	C	T	3.3	6	1	-	1	-	-	-	-	5	7	-	-	-	-	-

<i>Alt_{SB}_19</i>	0	A	1544	A	A	A	C	19	C	T	0.9	-	1	1	-	-	-	-	-	1	1	-	-	-	-	1
<i>Alt_{SB}_20</i>	7	G	1912	G	G	C	G	0	T	C	7.1	6	6	3	-	5	1	-	1	-	1	12	1	1	-	-
<i>Alt_{SB}_21</i>	7	G	1544	G	G	C	G	0	C	C	3.3	6	-	1	-	-	-	1	3	-	-	-	3	-	3	1
<i>Alt_{SB}_22</i>	0	A	1912	A	A	A	C	19	C	T	2.4	-	2	3	3	-	-	-	-	-	1	-	1	2	-	1
<i>Alt_{SB}_23</i>	0	A	1544	A	A	A	G	0	C	C	2.4	-	2	2	5	-	-	-	-	-	5	-	-	-	-	-
<i>Alt_{SB}_24</i>	7	G	1912	A	A	A	C	19	C	T	0.5	-	1	-	1	-	-	-	-	-	-	-	-	-	-	1
<i>Alt_{SB}_25</i>	0	G	1912	A	G	C	G	0	C	T	1.4	1	2	-	1	2	-	-	-	-	1	1	-	-	-	1
<i>Alt_{SB}_26</i>	0	G	1544	A	G	C	G	0	C	C	0.5	1	-	-	-	1	-	-	-	-	-	-	-	-	-	1
<i>Alt_{SB}_27</i>	7	G	1544	A	A	A	G	0	C	C	0.9	1	-	1	-	1	-	-	-	-	-	-	1	-	-	1
<i>Alt_{SB}_28</i>	7	G	1912	A	G	C	G	0	C	T	0.5	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-
<i>Alt_{SB}_29</i>	7	G	2280	A	G	C	G	0	C	C	5.2	11	-	-	-	-	-	3	7	-	-	-	1	-	10	-
<i>Alt_{SB}_30</i>	7	G	2280	G	G	C	G	0	T	C	0.5	-	1	-	1	-	-	-	-	-	-	-	-	1	-	-

^a S, sensível = CLR_{5d} < 30%; I, intermediária = 30% < CLR_{5d} < 80%; T=tolerante = CLR_{5d} > 80%.

^b G, guinea; C, caudatum; K, kafir; B, bicolor; D, durra; Gm, guinea margaritifera.

^c Q1, guinea do oeste da África e guinea margaritifera; Q2, caudatum da África e grupo de acessos transplantados caudatum e durra da região do Chad; Q3, linhagens melhoradas do Brasil e Estados Unidos; Q4, kafir do sul da África; Q5= Acessos durra do centro e leste da África e da Ásia e, acessos bicolor e caudatum da China; Q6, acessos guinea do sul da África e Ásia.

Os eventos de recombinação apresentados na Tabela 9 foram identificados com o auxílio do teste de quatro gametas (Hudson e Kaplan, 1985). O teste dos quatro gametas declara a ocorrência de recombinação em um determinado intervalo quando ele identifica os quatro gametas possíveis entre dois locos bialélicos. Entretanto, o teste não diferencia recombinação de mutação recorrente, o que pode superestimar o número de recombinações. Assumimos aqui que o haplótipo mais frequente é o haplótipo ancestral, mas na ausência de um *outgroup* característico (e.g. representante de outra espécie), não há condições de confirmar esta hipótese. Assim, a identificação dos haplótipos recombinantes foi baseada na presença de quatro gametas no intervalo indicado pelo teste, considerando-se *Alt_{SB}_2* como ancestral. Entretanto, neste conjunto de dados foi raramente possível a identificação de dois haplótipos parentais e dois haplótipos claramente recombinantes, considerando-se a permuta completa de toda a região posterior ao ponto de recombinação. Em muitos casos foram encontrados somente um dos haplótipos derivados do evento de recombinação putativo, enquanto os outros podem não ter sido amostrado neste conjunto de acessos (Tabela 10).

Um evento de recombinação envolvendo os haplótipos *Alt_{SB}_23* (representado por duas linhagens intermediárias e duas tolerantes ao Al, Tabela 10) e o haplótipo *Alt_{SB}_2*, que tem alta frequência de linhagens sensíveis, pode ter resultado no haplótipo *Alt_{SB}_27*, sendo representado por somente uma linhagem sensível e uma tolerante ao Al, SC566. Análises genéticas disponíveis para quatro linhagens tolerantes dos haplótipos *Alt_{SB}_23*, mostram que somente SC549 e IS26554 possuem alelos funcionais do gene *Alt_{SB}*. Como mostrado no capítulo 1, os padrões de diversidade genética obtidos com marcadores SSR indicam proximidade entre SC549 e SC566, classificados como guinea e caudatum, respectivamente. Adicionalmente, a raça guinea é simpátrica com todas as outras raças básicas e híbridos interracialis entre elas são observados (de Wet et al. 1972). Sugerindo que há possibilidade de hibridização entre os acessos portadores dos haplótipos *Alt_{SB}_23* e *Alt_{SB}_27*. Outro evento de recombinação envolvendo acessos portadores dos haplótipos *Alt_{SB}_22* e *Alt_{SB}_2* pode ter resultado no haplótipo *Alt_{SB}_24*, que neste conjunto de acessos está representado somente pelo acesso de tolerância intermediária, IS23100.

Tabela 10. Eventos de recombinação putativos envolvendo os haplótipos *Alt_{SB}_2* versus *Alt_{SB}_23*, *Alt_{SB}_2* versus *Alt_{SB}_22*, *Alt_{SB}_22* versus *Alt_{SB}_1*, *Alt_{SB}_29* versus *Alt_{SB}_5*. Posição inferida do evento de recombinação está indicada pela linha espessa.

Haplótipo	ORF6_160pb	ORF_200pb	MITE	SbMATE_I2_5947pb	SbMATE_I2_5985pb	SbMATE_I2_6083pb	SbMATE_I2_6094pb	OR9_12487pb	ORF10_24804pb	ORF10_25094pb	Tolerância		
											S	I	T
<i>Alt_{SB}_2</i>	7	G	1184	G	G	C	G	0	T	C	55	6	1
<i>Alt_{SB}_23</i>	0	A	1544	A	A	A	G	0	C	C	-	2	2
<i>Alt_{SB}_27</i>	7	G	1544	A	A	A	G	0	C	C	1	-	1
<i>Alt_{SB}_2</i>	7	G	1184	G	G	C	G	0	T	C	55	6	1
<i>Alt_{SB}_22</i>	0	A	1912	A	A	A	C	19	C	T	-	2	3
<i>Alt_{SB}_24</i>	7	G	1912	A	A	A	C	19	C	T	-	1	-
<i>Alt_{SB}_22</i>	0	A	1912	A	A	A	C	19	C	T	-	2	3
<i>Alt_{SB}_1</i>	7	G	456	A	G	C	G	0	C	C	48	1	-
<i>Alt_{SB}_15</i>	7	A	1184	A	A	A	C	19	C	T	1	-	-
<i>Alt_{SB}_26</i>	0	G	1544	A	G	C	G	0	C	C	1	-	-
<i>Alt_{SB}_29</i>	7	G	2280	A	G	C	G	0	C	C	11	-	-
<i>Alt_{SB}_5</i>	0	A	1184	A	A	C	G	0	C	T	4	-	-
<i>Alt_{SB}_26</i>	0	G	1544	A	G	C	G	0	C	C	1	-	-
<i>Alt_{SB}_13</i>	7	A	1184	A	A	C	G	0	C	T	1	-	-

3.5. Análise Filogenética da Região Alvo

A árvore de máxima parsimônia foi construída com os polimorfismos dos 30 haplótipos observados no loco *Alt_{SB}* visando elucidar os padrões de relacionamento entre esses haplótipos (Figura 10). A árvore com os haplótipos do loco *Alt_{SB}* revelou um agrupamento suportado por valores de *bootstrap* de 59% formado por basicamente por haplótipos raros (*Alt_{SB}_8*, *Alt_{SB}_15*, *Alt_{SB}_19* e *Alt_{SB}_24*) encontrados entre os acessos guinea do oeste e sul da África (Tabela 9). O *Alt_{SB}_22* presente neste grupo não é somente encontrado entre os acessos guinea do oeste e sul da África mas também entre as linhagens melhoradas do Brasil e Estados Unidos e entre os acessos kafir do sul da África (Tabela 9). Os demais agrupamentos de haplótipos foram suportados por valores baixos de *bootstrap*, possivelmente devido ao alto número de haplótipos recombinantes que pode ter dificultado a formação de grupos. Estes agrupamentos também não

apresentaram padrões coerentes com estrutura de população. Entretanto, a localização do haplótipo comum entre os acessos guinea do oeste da África, *Alt_{SB}_23* (Q1, Tabela 9) e do haplótipo comum entre os acessos caudatum da África e grupo de acessos transplantados caudatum e durra da região do Chad, *Alt_{SB}_20* (Q2, Tabela 9), não sugere relacionamento entre estes haplótipos (Figura 10), o que pode ter sido a causa dos diferentes resultados obtidos com a associação quando a importância destas raças foram avaliadas.

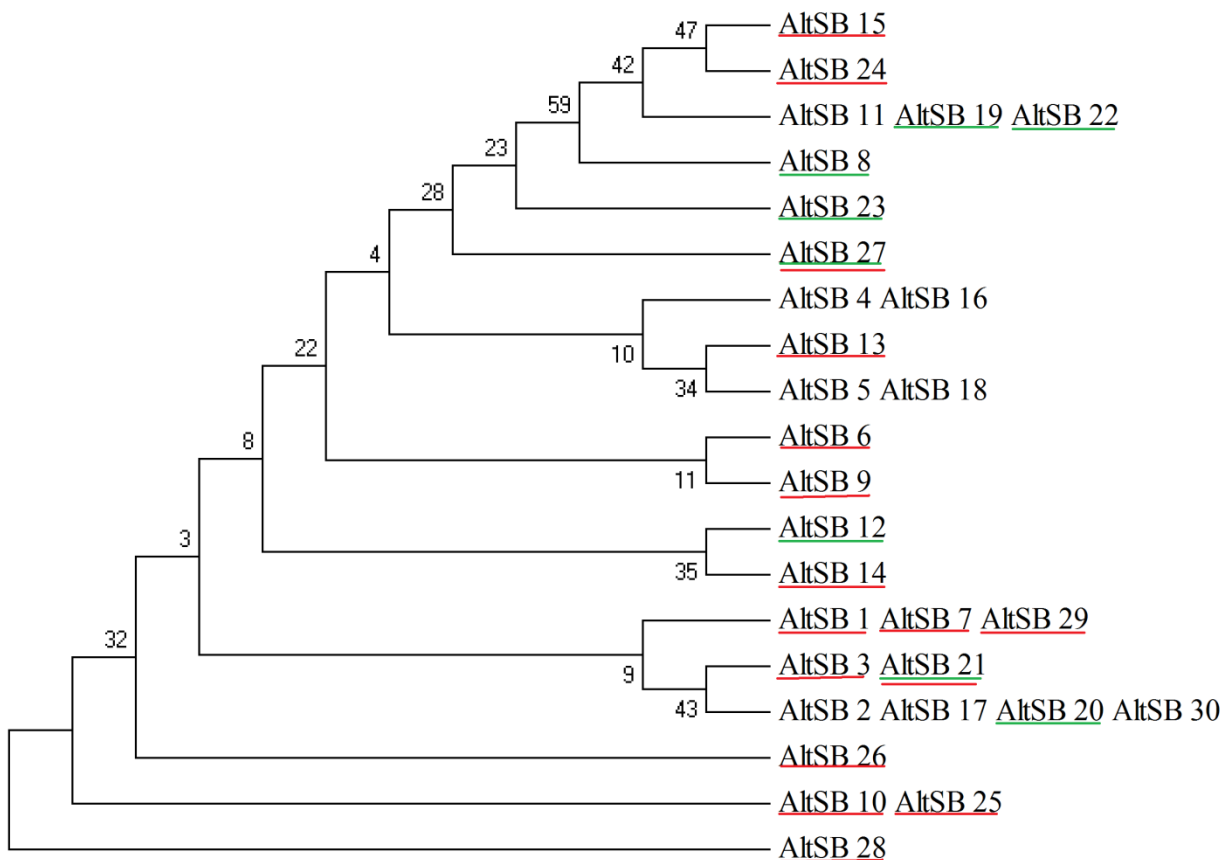


Figura 10. Árvore de máxima parsimonia baseada nos dados de sequência dos 30 haplótipos do loco *Alt_{SB}*. Hapótipos sublinhados com a cor verde indicam a presença de acessos tolerantes ao Al e os sublinhados em vermelho indicam haplótipos recombinantes. Valores de 1000 *bootstrap* estão apresentados em cada ramo.

3.6. Estratégias de Seleção

Os resultados da análise de associação irão nos permitir desenvolver marcadores gene

específicos que serão aplicados na seleção assistida por marcadores (SAM) visando à introgressão desta característica em linhagens elite, como também para o screening de germoplasma visando à busca de novas fontes de tolerância ao Al. A Tabela 11 mostra os genótipos dos quatorze acessos tolerantes do painel ($CLR_{5d} < 80\%$) para os locos associados com tolerância ao Al e a frequência do alelo favorável no painel completo.

Tabela 11. Genótipo dos quatorze acessos tolerantes, $CLR_{5d} > 80\%$, para os locos associados com tolerância ao Al. Estão indicados os alelos de cada loco com o alelo favorável indicado em negrito, seguido pela sua frequência no painel completo. O sinal + e células sombreadas indicam a presença do alelo favorável entre os acessos tolerantes e o sinal – indica a presença do alelo alternativo.

Loco	Alelo	Frequência	Acessos Tolerantes														Total
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	
ORF6_160pb	0/7	0.20	+	+	-	-	-	-	-	+	+	+	+	+	-	+	8
ORF6_200pb	A/G	0.18	+	+	-	-	-	-	-	+	+	+	+	+	-	+	8
MITE_1184	+/-	0.44	+		+											+	3
MITE_1912pb	+/-	0.12	-	-	-	+	+	+	-	+	+	+	-	-	-	-	6
<i>SbMATE_I2_6083pb</i>	A/C	0.10	-	+	-	-	-	-	-	+	+	+	+	+	+	+	8
<i>SbMATE_I2_6094pb</i>	C/G	0.06	-	+	-	-	-	-	-	+	+	+	-	-	-	+	5
ORF9_12487pb	19/0	0.06	-	+	-	-	-	-	-	+	+	+	-	-	-	+	5
ORF10_24804pb	C/T	0.60	+	+	-	-	-	-	+	+	+	+	+	+	+	+	10

Os alelos favoráveis nos locos ORF6_160pb, ORF6_200pb, MITE_1184 e ORF10_24804pb estão em frequência superior à dos acessos tolerantes que é de 6%, indicando que a utilização destes locos como marcadores poderá resultar também na seleção de acessos intermediários e sensíveis. Entretanto, os alelos favoráveis para os locos MITE_1912pb, *SbMATE_I2_6083pb*, *SbMATE_I2_6094pb* e ORF9_12487pb estão em frequência baixa no conjunto de acessos e poderão resultar na seleção mais precisa de acessos tolerantes. Dentre estes os locos MITE_1912pb e *SbMATE_I2_6083pb* parecem mais interessantes, uma vez que eles recuperam seis e oito acessos tolerantes, respectivamente. Enquanto que os alelos favoráveis dos locos *SbMATE_I2_6094pb* e ORF9_12487pb recuperam cada um somente cinco acessos tolerantes. Finalmente, MITE_1912pb e *SbMATE_I2_6083pb* recuperam juntos onze acessos tolerantes. Não foi possível a recuperação de dois acessos altamente tolerantes, que possivelmente possuem outros locos de tolerância distintos do *Alt_{SB}*.

4. Discussão

Para entender a diferenciação genética, aspectos de adaptação, conservação e usos potenciais do germoplasma de sorgo, muitos estudos de diversidade genética têm sido conduzidos que mostraram uma organização da diversidade da espécie compatível com padrões raciais e geográficos (Aldrich et al. 1992; Tao et al. 1993; Deu et al. 1994, 2006; Cui et al. 1995; de Oliveira et al. 1996; Menkir et al. 1997; Djè et al. 2000; Barnaud et al. 2007). Algumas características fenotípicas podem ocorrer com maior frequência em determinados compartimentos da diversidade global da espécie, como é o caso da tolerância ao Al, em que a ocorrência de acessos tolerantes é mais frequente nas raças *caudatum* e *guinea* (Reddy et al. 2006; capítulo 1). Diante desse cenário, a desconsideração do fator estrutura de população pode nos levar a declarar um loco como associado à tolerância ao Al devido ao desequilíbrio de ligação causado por estruturação populacional e não ligação física a QTNs de tolerância ao Al. Ou seja, a chance de se cometer Erro Tipo I é aumentada devido à ocorrência conjunta de estrutura de população e de diferenças fenotípicas para tolerância ao Al entre subpopulações. Nossos resultados mostram que o modelo linear misto, com seis subpopulações e a matriz de relacionamento familiar baseada no número de alelos compartilhados fornece um controle adequado do Erro Tipo I, resultando em um menor número de associações falso positivas. Neste conjunto de dados, o controle da estrutura de população apresentou maior importância do que o controle de parentesco, como observado pela contribuição isolada de ambas as matrizes Q e K no controle das falsas associações. Entretanto, o controle exercido por Q+K foi ainda melhor, indicando que ambos devem ser considerados no modelo. A contribuição do fator estrutura de população não parece ter sido crítica para o controle do Erro Tipo I em um conjunto de 377 linhagens de sorgo representado por 60% de linhagens do programa de conversão, (linhagens SC, *sorghum conversion*, Casa et al. 2008). É possível que esse resultado guarde relação com a alta representação das linhagens do programa de conversão no painel. Linhagens SC são caracterizadas por apresentarem porte pequeno e florescimento independente do comprimento do dia, e foram desenvolvidas para permitir o cultivo de sorgo em regiões temperadas. Elas foram desenvolvidas de linhagens altas e sensíveis ao fotoperíodo por meio da introgressão de alelos de nanismo e insensibilidade ao fotoperíodo de um doador comum (Stephens et al. 1967). Considerando o limitado número de retrocruzamentos realizados, regiões relativamente extensas podem ter sido introgridas por arraste genético com as regiões

contendo locos relacionados ao processo de conversão. Isso pode ter resultado em estrutura populacional de origem recente ou relacionamento familiar entre indivíduos, com consequente maior importância da correção para parentesco (K) no controle de falsos positivos nos testes de associações das linhagens SC.

A queda do desequilíbrio de ligação para valores próximos a zero dentro de 5000pb na região do loco *Alt_{SB}* difere dos valores obtidos para o gene *Opaco 2* (O2) onde foi observado desequilíbrio de ligação completo dentro de uma região de 4250pb no loco sob ausência de seleção e recombinação (de Alencar Figueiredo et al. 2008). Desequilíbrio de ligação completo com raras evidências de recombinação foi também observado dentro de uma região de 1380pb do gene *Amylose extender1* (*Ae1*), que sofreu pressão seletiva. Entretanto, para outro gene onde houve efeito seletivo, *Waxy* (*Wx*), a queda do desequilíbrio de ligação foi provavelmente causada pela intensa recombinação dentro de uma região de 244pb que separa o intron 6 do intron 7 deste gene (de Alencar Figueiredo et al. 2008). Existem várias causas para o declínio mais rápido do desequilíbrio de ligação na região do loco *Alt_{SB}*. A não significância do teste D de Tajima conduzido em regiões que acomodam os polimorfismos associados sugere que a evolução do loco *Alt_{SB}* deu-se segundo a teoria neutra, sem influência marcante de seleção. Adicionalmente, a região estudada está localizada na porção distal do cromossomo 3 (Magalhaes et al. 2004) e, o sequenciamento completo do genoma do sorgo, juntamente ao seu alinhamento aos mapas genéticos, indicou que eventos de recombinação são confinados à aproximadamente um terço do genoma da espécie, sendo mais frequentes nas porções terminais dos cromossomos (Paterson et al. 2009). De fato, o gene *Alt_{SB}* foi clonado por meio da técnica de clonagem posicional, que permitiu a definição de uma região alvo de apenas 24,6kpb que continha o gene. Isso só foi possível devido à alta taxa de recombinação na região, que produziu uma razão entre distância física e genética favorável ao mapeamento de alta resolução.

A extensão física do desequilíbrio de ligação determina a estratégia de mapeamento e seleção a serem empregados. Se o desequilíbrio é muito extenso, como tem sido observado em espécies autógamas como o sorgo (Hamblin et al. 2005) e arroz (Garris et al. 2003), a procura por associações no genoma é factível uma vez que o número de marcadores necessários para a sua cobertura pode tornar-se economicamente viável. Entretanto, blocos de ligação na região do QTN limitam a identificação dos polimorfismos causativos (Hamblin et al. 2007) ou podem

resultar em associações significativas com blocos gênicos, sem a individualização do gene responsável pelo fenótipo. Entretanto, em uma situação em que há queda rápida do desequilíbrio de ligação como em espécies de fecundação cruzada como o milho, em que o declínio para algumas regiões ocorre dentro de 100pb a 7000pb (Remington et al. 2001), a busca de associações no genoma completo da espécie torna-se economicamente proibitiva, mas a busca de polimorfismos causativos torna-se possível. A exemplo da recente identificação de quatro polimorfismos do loco *lycopene epsilon cyclase (lcyE)* identificados como responsáveis por 58% da variação presente na via de produção de vitamina A (Harjes et al. 2008)

Os dez polimorfismos associados explicaram, relativamente, uma pequena parte da variação fenotípica da característica, variando de 2,6 a 13,5%. Esses valores são baixos quando comparados aos 80% alcançados com marcadores flanqueando o loco *Alt_{SB}* (Caniato et al. 2007) em uma população biparental. Isto pode ser explicado pela natureza biparental da população de mapeamento, que concentra grande parte da variação fenotípica em um único loco. A análise de associação aqui apresentada foi baseada em um conjunto de acessos diversos, com grande variabilidade fenotípica para a tolerância ao Al. É possível também que heterogeneidade genética esteja ocorrendo na região do loco *Alt_{SB}*, contribuindo para a baixa variação explicada pelos alelos dos locos associados. Uma indicação disso é o fato do haplótipo comum entre os acessos da raça caudatum, *Alt_{SB}_20*, ser pouco relacionado aos haplótipos mais frequentes entre os acessos tolerantes da raça guinea, *Alt_{SB}_22* e *Alt_{SB}_23*. Heterogeneidade genética tem sido reportada para outras espécies autógamas como *Arabidopsis* e arroz. Associação para tempo de florescimento em *Arabidopsis* identificou que os haplótipos associados com o florescimento precoce no loco *FRIGIDA*, comumente encontrado entre os acessos do centro da Ásia, são diferentes daqueles encontrados nos demais acessos (Hagenblad et al. 2004). Similarmente, diferentes haplótipos associados com a reação a resistência a *Xanthomonas oryzae* pv. *oryzae* no loco *xa5* de arroz, foram observados nos acessos classificados como Indica e Aus-boro (Garris et al. 2003). Entretanto, é também possível que a tolerância da raça caudatum venha, pelo menos em parte, de outros locos de tolerância distintos do loco *Alt_{SB}*, como sugerido pela falta de ligação de marcadores ligados ao loco *Alt_{SB}* em populações de mapeamento derivadas de membros da raça caudatum, como SC112 (Caniato et al. 2007).

Dentre os polimorfismos significativamente associados com a tolerância ao Al, o SNP *SbMATE_I2_6083pb* explicou uma maior parte da variação fenotípica que os outros polimorfismos, particularmente para o CP1, que é um índice global de tolerância, e para resposta indutiva. A resposta indutiva em sorgo é caracterizada por um aumento expressivo do crescimento radicular entre 3 e 5 dias de exposição ao Al, período em que ocorre, de forma paralela, um aumento dos níveis de expressão do gene *Alt_{SB}*. Portanto, o fato do loco *SbMATE_I2_6083pb* ter explicado uma maior porção da variância fenotípica especificamente para CLR aos 5 dias e para o índice b, que mede a indução do crescimento radicular entre 3 e 5 dias, sugere um efeito individualizado desse loco na resposta indutiva. Esse SNP do loco *SbMATE_I2_6083pb* está localizado no segundo intron do gene *SbMATE*. Tem sido descrito que introns podem afetar a expressão gênica de duas maneiras, aumentando a síntese de mRNA através de um ativador (*enhancer*) ou como um promotor alternativo localizado dentro do intron (Rose 2008). Entretanto, o papel deste polimorfismo como mediador da expressão gênica precisará ser validado posteriormente.

A estimativa de diversidade nucleotídica na região completa do loco *Alt_{SB}*, $\delta=1,10/\text{kbp}$, foi baixa em relação a estimativa de $2.05/\text{kbp}$ obtida em seis genes candidatos envolvidos com biosíntese de amido e regulação do conteúdo de proteína no endosperma dos grãos de sorgo (de Alencar Figueiredo et al. 2008). Desvio do modelo neutro de evolução foi observado para o segundo exon do *SbMATE*, com valor de D negativo revelando grande quantidade de alelos raros, que podem indicar padrão de seleção purificadora, ou seja, a seleção deve estar atuando no sentido de eliminar as mutações que surgem na região codante que poderão comprometer a funcionalidade da proteína. O MITE apresenta grande importância na diversidade haplotípica observada, pois quando os alelos do MITE foram removidos, o número de haplótipos foi reduzido para 19. Isto pode ser devido a natureza repetitiva deste tipo de loco que apresenta taxa de mutação diferenciada dos demais (Wessler et al. 1995). Adicionalmente, os eventos de recombinação também apresentaram papel importante para a geração de diversidade haplotípica recente, devido ao grande número de haplótipos raros possivelmente associados com eventos de recombinação.

Nossos resultados revelaram associação significativa entre tolerância ao Al e dez polimorfismo na região de 24,6kpb onde está localizado o loco *Alt_{SB}*, identificados em um conjunto de acessos altamente diversos. Dentre os locos associados o *SbMATE_6083pb*

explicou a maior da variação para resposta indutiva, um importante componente da tolerância ao Al. A baixa frequência deste loco o faz promissor para a mineração de alelos de tolerância condicionados pelo loco *Alt_{SB}*, uma vez que a sua frequência é tão baixa quanto a da característica. Entretanto, a dificuldade de se identificar o polimorfismo causativo é evidente diante do desequilíbrio de ligação significativo apresentado pelos locos associados, não sendo possível distinguir se a associação entre um dado SNP e o fenótipo foi resultado da identificação do QTN ou foi resultado de um SNP em desequilíbrio de ligação com o QTN. Nossos resultados também revelam completa queda do desequilíbrio após 5000pb, causada pelo conjunto de fatores: recombinação, ausência de seleção na região e a localização da região estudada na porção distal do cromossomo 3. Finalmente, nossos resultados e os literatura (Yu et al. 2006, Casa et al. 2008) indicam a importância do controle do Erro Tipo I nos estudos de associação, visando controlar estatisticamente o desequilíbrio de ligação causados por fatores outros que a ligação física. Como observado, a ausência de controle de estrutura de população e parentesco resultam em maior número de locos associados e na superestimação da variância fenotípica explicada pelos locos associados. Espera-se portanto que os resultados e as estratégias aqui apresentadas possam ser úteis para a identificação e seleção de fontes de tolerância ao Al e sua transferência para germoplasma elite em programas de melhoramento de sorgo para solos ácidos.

5. Anexo

Anexo 1: Haplótipos identificados na região de 24,6kpb do loco *Alt_{SB}*. As posições dos intervalos de recombinação foram estimadas pelo teste dos quatro gametas (Hudson e Kaplan, 1985). A atribuição dos eventos de recombinação putativos aos haplótipos foi feita assumindo-se que o haplótipo ancestral corresponde a *Alt_{SB}_2* (portanto, todos os haplótipos recombinantes foram definidos com base no haplótipo *Alt_{SB}_2*). As barras finas identificam a recombinação entre os alelos dos locos ORF6_160pb e ORF6_200pb. As barras grossas indicam os intervalos de recombinação entre os alelos dos locos ORF6_200pb e *SbMATE*_5947pb e entre os alelos dos locos *SbMATE*_5985pb e ORF10_24804pb. À direita estão identificadas a linhagens descritas no texto

Haplótipo	ORF6_160pb	ORF6_200pb	MITE	SbMATE_5947pb	SbMATE_5985pb	SbMATE_6083pb	SbMATE_6094pb	OR9_12487pb	ORF10_24804pb	ORF10_25094pb	Acessos
<i>Alt_{SB}_1</i>	7	G	456	A	G	C	G	0	C	C	IS29691, BR007B
<i>Alt_{SB}_2</i>	7	G	1184	G	G	C	G	0	T	C	
<i>Alt_{SB}_3</i>	7	G	1184	G	G	C	G	0	C	C	
<i>Alt_{SB}_4</i>	0	A	1184	A	A	C	G	0	C	C	
<i>Alt_{SB}_5</i>	0	A	1184	A	A	C	G	0	C	T	
<i>Alt_{SB}_6</i>	0	G	1184	A	A	C	G	0	C	C	
<i>Alt_{SB}_7</i>	7	G	1184	A	G	C	G	0	C	C	
<i>Alt_{SB}_8</i>	0	A	1184	A	A	C	C	19	C	C	
<i>Alt_{SB}_9</i>	7	G	1184	A	A	C	G	0	T	C	
<i>Alt_{SB}_10</i>	0	G	1184	A	G	C	G	0	C	T	
<i>Alt_{SB}_11</i>	0	A	1184	A	A	A	C	19	C	T	IS25077
<i>Alt_{SB}_12</i>	0	A	1184	A	G	C	G	0	C	C	
<i>Alt_{SB}_13</i>	7	A	1184	A	A	C	G	0	C	T	
<i>Alt_{SB}_14</i>	0	A	1184	G	G	C	G	0	C	C	
<i>Alt_{SB}_15</i>	7	A	1184	A	A	A	C	19	C	T	
<i>Alt_{SB}_16</i>	0	A	1544	A	A	C	G	0	C	C	
<i>Alt_{SB}_17</i>	7	G	1544	G	G	C	G	0	T	C	
<i>Alt_{SB}_18</i>	0	A	1544	A	A	C	G	0	C	T	
<i>Alt_{SB}_19</i>	0	A	1544	A	A	A	C	19	C	T	
<i>Alt_{SB}_20</i>	7	G	1912	G	G	C	G	0	T	C	IS10801, IS21849 SC175, SC112, 3DX, 5DX, 9DX BR012R, IS23142 SC283, CMSXS225, IS14351 SC549, IS26554 IS23100
<i>Alt_{SB}_21</i>	7	G	1544	G	G	C	G	0	C	C	
<i>Alt_{SB}_22</i>	0	A	1912	A	A	A	C	19	C	T	
<i>Alt_{SB}_23</i>	0	A	1544	A	A	A	G	0	C	C	
<i>Alt_{SB}_24</i>	7	G	1912	A	A	A	C	19	C	T	
<i>Alt_{SB}_25</i>	0	G	1912	A	G	C	G	0	C	T	
<i>Alt_{SB}_26</i>	0	G	1544	A	G	C	G	0	C	C	
<i>Alt_{SB}_27</i>	7	G	1544	A	A	A	G	0	C	C	
<i>Alt_{SB}_28</i>	7	G	1912	A	G	C	G	0	C	T	
<i>Alt_{SB}_29</i>	7	G	2280	A	G	C	G	0	C	C	
<i>Alt_{SB}_30</i>	7	G	2280	G	G	C	G	0	T	C	SC566

6. Literatura Citada

ALDRICH, P.R.; DOEBLEY, J.; SCHERTZ, K.F.; STEC, A. Patterns of allozyme variation in cultivated and wild sorghum bicolor. **Theoretical and Applied Genetics**, Berlin, v.114, p. 451-460, 1992.

BARNAUD, A.; DEU, M.; GARINE, E.; MCKEY, D.; JOLY, H.I.; Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. **Theoretical and Applied Genetics**, Berlin, v.114, p. 237-248, 2007.

BUREAU, B.E; WESSLER, S.R. Tourist: A large family of small inverted repeat elements frequently associated with maize genes. **Plant Cell**, v.4, p. 1283-1294, 1994.

CANIATO, F.F.; GUIMARÃES, C.T.; SCHAFFERT, R.E.; ALVES, V.M.; KOCHIAN L.V.; BORÉM, A.; KLEIN, P.E.; MAGALHAES J.V. Genetic Diversity for Al Tolerance in Sorghum. **Theoretical and Applied Genetics**, Berlin, v.114, p. 863-876, 2007.

JIN.L.; CHAKRABORTY, R. Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. **Molecular Biology and Evolution**, v. 11, p.120-127, 1994.

CASA, A.M.; PRESSOIR, G.; BROWN, P.J.; MITCHELL, S.E.; ROONEY, W.L.; TUINSTRA.M.R.; FRANKS,C.D.; KRESOVICH, S. Community Resources and Strategies for Association Mapping in Sorghum. **Crop Science**, v. 48, p. 30-40, 2008.

CUI,Y.X.; XU, G.W.; MAGILL, C.W.; SCHERTZ, K.F. RFLP-based assay of Sorghum bicolor (L) Moench genetic diversity. **Theoretical and Applied Genetics**, Berlin, v. 90, p. 787-796, 1995.

de ALENCAR FIGUEIREDO, L.F.; CALATAYUD, C.; DUPUITS, C.; BILLOT, C.; RAMI, J.-F.; BRUNEL, D.; PERRIER, X.; COURTOIS, M.; DEU, M.; GLASZMANN, J.-C. Phylogeographic evidence of crop neodiversity in sorghum. **Genetics**, Maryland, v. 179, p. 997-1008, 2008.

de OLIVEIRA, A.C.; RICHTER, T.; BENNETZEN, J.L. Regional and racial specificities in sorghum germplasm assessed with DNA markers. **Genome**, v. 39, p. 579-587, 1996.

de SILVA, H.N.; BALL, R.D. Linkage disequilibrium mapping concepts. In: ORAGUZIE, N.C.; RIKKERINK, E.H.A.; GARDINER, S.E.; DE SILVA, H.N. (Eds.) **Association Mapping in Plants**. New York, US: Springer, 2007. p. 103-132.

de Wet, J.M.J.; Harlan, J.R.; Kurmarohita, B. Origin and evolution of Guinea sorghums. **East African Agricultural and Forestry Journal**, v. 38, p. 114-119, 1972.

DJÈ, Y.; HEUERTZ, M.; LEFÈBVRE, C.; VEKEMANS, X. Assessment of genetic diversity within and among germoplasm accessions in cultivated sorghum using microsatellite markers. **Theoretical and Applied Genetics**, Berlin, v.100, p. 918-925, 2000.

DEU, M.; GONZALEZ-DE-LEON, D.; GLASZMANN, J.C.; DEGREMONT, I. CHANTEREAU, J.; LANAUD, C.; HAMONN, P. RFLP diversity in cultivated sorghum in relation to racial differentiation. **Theoretical and Applied Genetics**, Berlin, v.88, p. 838-844, 1994.

DEU, M.; RATTUNDE, F.; CHANTEREAU, J. A global view of genetic diversity in cultivated sorghums using a core collection. **Genome**, Ottawa, v. 49, p.168-180, 2006

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular Ecology**. v. 14, p. 2611-2620, 2005.

FLINT-GARCIA, S.A.; THORNSBERRY, M.J.; BUCKLER, E. Structure of linkage

disequilibrium in plants. **Annual Review of Plant Biology**. v. 54, p. 357-374, 2003.

GARRIS, A.J.; McCOUCH, S.R.; KRESOVICH, S. Population structure and its effects on haplotypes diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). **Genetics**, Maryland, v. 165, p. 759-769, 2003.

HAGENBLAD, J.; TANG, C.; MOLITOR, J.; WERNER, J.; ZHAO, K.; ZHENG, H.; MARJORAM, P.; WEIGEL, D.; NORDBORG. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. **Genetics**, Maryland, v. 168, p. 1627-1638, 2004.

HARJES, C.E.; ROCHEFORE, T.R.; BAI, L. BRUTNELL, T.P.; KANDIANIS, C.B.; SOWINSKI, S.G.; STAPLETON, A.E.; VALLABHANENI, R.; WILLIAMS, M.; WURTZEL, E.T.; YAN, J.; BUCKLER, E.S. Natural genetic variation in Lycopene epsilon cyclase tapped for maize biofortification. **Science**, v. 319, p. 330-333, 2008.

HAMBLIN, M.T.; FERNANDEZ, M.G.S.; CASA, A.M.; MITCHELL, S.E.; PATERSON, A.H.; KRESOVICH, S. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. **Genetics**, Maryland, v. 171, p. 1247-1256, 2005.

HAMBLIN, M.T.; FERNANDEZ, M.G.S.; TUINSTRA, M.R.; ROONEY, W.R.; KRESOVICH, S. Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. **Crop Science**, v. 47, n.2, p. s125-134, 2007.

HILL, W.G.; ROBERTSON, A. Effects of inbreeding at loci with heterozygote advantage. **Genetics**, Maryland, v. 60, p. 615-628, 1968.

HUDSON, R.R; KAPLAN, N.L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. **Genetics**, Maryland, v. 111, p. 147-163, 1985.

KOCHIAN, L.V.; HOEKENGA, O.A.; PIÑEROS, M.A. How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. **Annual Review of Plant Biology**, v. 55, p. 459-493, 2004.

KUMAR, S.; TAMURA, K.; JAKOBSEN, I. B.; NEI, M. MEGA2: Molecular Evolutionary Genetics Analysis software. **Bioinformatics**, v.17, p. 1244-1245, 2001.

LEWONTIN, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. **Genetics**, Maryland, v. 49, p.49-67, 1964.

LIU, K.; MUSE, S.V. PowerMarker: integrated analysis environment for genetic marker data. **Bioinformatics**, v. 21, p. 2128-2129, 2005.

MAGALHAES, J.V.; GARVIN, D.F.; WANG, Y.; SORRELLS, M.E.; KLEIN, P.E.; SCHAFFERT, R.E.; LI, L.; KOCHIAN, L.V. Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the Poaceae. **Genetics**, Maryland, v. 167, p. 1905-1914, 2004.

MAGALHAES, J.V.; LIU, J.; GUIMARÃES, C.T.; LANA, U.G.P.; ALVES, V.M.C.; WANG, Y.H.; SCHAFFERT, R.E.; HOEKENGA, O.A.; PIÑEROS, M.A. SHAFF, J.E.; KLEIN, P.E.; CARNEIRO, N.P.; COELHO, C.M.; TRICK, A.N.; KOCHIAN, L.V. A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. **Nature Genetics**, v. 39, p. 1156-1161, 2007.

MENKIR, A.; GOLDSBROUGH, P.; EJETA, G. RAPD based assessment of genetic diversity in cultivated races of sorghum. **Crop Science**, v. 37, p. 564-569, 1997.

PARKER, D. R.; NORVELL, W. A.; CHANEY, R. L. GEOCHEM-PC: a chemical speciation program for IBM and compatible computers. In: LOEPPERT, R. H., et al. (Ed.) **Chemical**

equilibrium and reaction models. Madison: Soil Science Society of America, 1995. p. 253-269.

PATERSON, A.H.; BOWERS, J.E.; BRUGGMANN, R.; DUBCHAK, I.; GRIMWOOD, J.; GUNDLACH, H.; HABERER, G.; HELLSTEN, U.; MITROS, T.; POLIAKOV, A.; SCHMUTZ, J.; SPANNAGL, M.; TANG, H.; WANG, X.; WICKER, T.; BHARTI, A.K.; CHAPMAN, J.; ALEX FELTUS, F. GOWIK, U.; GRIGORIEV, I.V.; LYONS, E. MAHER, C.A.; MARTIS, M.; NARECHANIA, A.; OTILLAR, R.P.; PENNING, B.W.; SALAMOV, A.A.; WANG, Y.; ZHANG, L.; CARPITA, N.C.; FREELING, M.; GINGLE, A.R.; HASH, C.T.; KELLER, B.; KLEIN, P.; KRESOVICH, S.; MAcCANN, M.C.; MING, R.; PETERSEN, D.G.; RAHMAN, M.U.; WARE, D.; WESTHOFF, P.; MAYERS, K.F.X.; MESSING, J.; ROKHSAR, D.S. The Sorghum bicolor genome and the diversification of grasses. **Nature**, v. 457, p. 551-556, 2009.

PRITCHARD, J.K. STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Maryland, v. 155, p. 945-959, 2000.

PRITCHARD, J.K. Deconstructing maize population structure. **Nature Genetics**. v. 28, p. 203-204, 2001.

REDDY, B.V.S.; RAMESH, S.; REDDY, P.S. Sorghum Genetics Resources, Cytogenetics and Improvement. In: SINGH,R.J.; JAUHAR, P.P. (Eds.) **Genetic Resources, Chromosome Engineering and Crop Improvement: Vol 2: Cereals**. Boca Raton: Taylor & Francis Group, 2006. p. 309-363.

REMINGTON, D.L.; THORNSBERRY, J.M.; MATSUOKA, Y.; WILSON, L.M., WHITT, S.R.; DOEBLEY, J.F.; KRESOVICH, S.; GOODMAN, M.M.; BUCKLER, E.S. Structure of linkage disequilibrium and phenotypic association in maize genome. **Proceedings of the National Academy of Science of USA**. v. 25, p. 11479-11484, 2001.

ROSE, A.B. Intron-Mediated Regulation of Gene Expression. In: Reddy, A.S.N.; Golovkin, M (Eds.) **Nuclear pre-mRNA Processing in Plants: Current Topics in Microbiology and Immunology**. Berlin: Springer-Verlag, 2008. p. 277-290.

SAGHAI-MAROOF, M. A.; SOLIMAN, K. A.; JORGENSEN, R. A.; ALLARD, R. W. Ribosomal DNA spacer length polymorphism in barley: Mendelian inheritance, chromosomal location and population dynamics. **Proceedings of the National Academy of Science of USA**, Washington, v. 81, p. 8014-8018, 1984.

STEPHENS, J.C.; MILLER, F.R.; ROSENOW, D.T. Conversion of alien sorghums to early combine genotypes. **Crop Science**, v.7, p.396-396, 1967.

TAJIMA, F. Statistical Method for testing the neutral mutation hypothesis by DNA polymorphism, **Genetics**, Maryland, v. 123, p. 585-595, 1989.

TAO, Y.; MANNERS, J.M.; LUDLOW, M.M.; HENZELL, R.G. DNA polymorphisms in grain sorghum (*Sorghum bicolor* (L) Moench). **Theoretical and Applied Genetics**, Berlin, v. 86, p. 679-688, 1993.

WATTERSON, G.A. On the number of segregating sites in genetical models without recombination. **Theoretical Population Biology**, v. 7, p. 256-276, 1975.

WEBER, A.L.; BRIGGS, W.H.; RUCKER, J.; BALTAZAR, B.M.; SÁNCHEZ-GONZALEZ, J.J.; FENG, P.; BUCKLER, E.S.; DOEBLEY, J.F. The genetic architecture of complex traits in Teosinte (*Zea mays ssp. Parviglumis*): New evidence from association mapping. **Genetics**, Maryland, v. 180, p. 1221-1232, 2008.

WESSLER, S.R.; BUREAU, B.E.; WHITE, S.E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. **Current Biology**, v. 5, p. 814-821, 1995.

WRIGHT. S. The genetical structure of population. **Annals of Eugenics**, v. 15, p. 323-354, 1951

YU, J.; PRESSOIR, G.; BRIGGS, W.H.; BI, I.V.; YAMASAKI, M.; DOEBLEY, J.F.; McMULLEN, M.D.; GAUT, B.S.; NIELSEN, D.M.; HOLLAND, J.B.; KRESOVICH, S.; BUCKLER, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**. v. 38, p. 203-208, 2006.