

# Association Mapping Provides Insights into the Origin and the Fine Structure of the Sorghum Aluminum Tolerance Locus, *Alt<sub>SB</sub>*

Fernanda F. Caniato<sup>1</sup>, Martha T. Hamblin<sup>2</sup>, Claudia T. Guimaraes<sup>1</sup>, Zhiwu Zhang<sup>2</sup>, Robert E. Schaffert<sup>1</sup>, Leon V. Kochian<sup>3</sup>, Jurandir V. Magalhaes<sup>1\*</sup>

**1** Embrapa Maize and Sorghum, Sete Lagoas, Minas Gerais, Brazil, **2** Institute for Genomic Diversity, Cornell University, Ithaca, New York, United States of America, **3** Robert W. Holley Center for Agriculture and Health, U.S. Department of Agriculture – Agricultural Research Service, Cornell University, Ithaca, New York, United States of America

## Abstract

Root damage caused by aluminum (Al) toxicity is a major cause of grain yield reduction on acid soils, which are prevalent in tropical and subtropical regions of the world where food security is most tenuous. In sorghum, Al tolerance is conferred by *SbMATE*, an Al-activated root citrate efflux transporter that underlies the major Al tolerance locus, *Alt<sub>SB</sub>*, on sorghum chromosome 3. We used association mapping to gain insights into the origin and evolution of Al tolerance in sorghum and to detect functional variants amenable to allele mining applications. Linkage disequilibrium across the *Alt<sub>SB</sub>* locus decreased much faster than in previous reports in sorghum, and reached basal levels at approximately 1000 bp. Accordingly, intra-locus recombination events were found to be extensive. SNPs and indels highly associated with Al tolerance showed a narrow frequency range, between 0.06 and 0.1, suggesting a rather recent origin of Al tolerance mutations within *Alt<sub>SB</sub>*. A haplotype network analysis suggested a single geographic and racial origin of causative mutations in primordial guinea domesticates in West Africa. Al tolerance assessment in accessions harboring recombinant haplotypes suggests that causative polymorphisms are localized to a ~6 kb region including intronic polymorphisms and a transposon (MITE) insertion, whose size variation has been shown to be positively correlated with Al tolerance. The SNP with the strongest association signal, located in the second *SbMATE* intron, recovers 9 of the 14 highly Al tolerant accessions and 80% of all the Al tolerant and intermediately tolerant accessions in the association panel. Our results also demonstrate the pivotal importance of knowledge on the origin and evolution of Al tolerance mutations in molecular breeding applications. Allele mining strategies based on associated loci are expected to lead to the efficient identification, in diverse sorghum germplasm, of Al tolerant accessions able maintain grain yields under Al toxicity.

**Citation:** Caniato FF, Hamblin MT, Guimaraes CT, Zhang Z, Schaffert RE, et al. (2014) Association Mapping Provides Insights into the Origin and the Fine Structure of the Sorghum Aluminum Tolerance Locus, *Alt<sub>SB</sub>*. PLoS ONE 9(1): e87438. doi:10.1371/journal.pone.0087438

**Editor:** James C. Nelson, Kansas State University, United States of America

**Received:** September 12, 2013; **Accepted:** December 24, 2013; **Published:** January 30, 2014

**Copyright:** © 2014 Caniato et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors acknowledge funding from the CGIAR Generation Challenge Program (project G3007.04), the Embrapa Macroprogram and the McKnight Foundation Collaborative Crop Research Program. The authors are also thankful to the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for a post-doctoral fellowship granted to FFC and to The National Council for Scientific and Technological Development (CNPq) for support to JVM and CTG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: jurandir.magalhaes@embrapa.br

## Introduction

Among the various abiotic stresses that limit sorghum production, aluminum (Al) toxicity has been identified as one of the main breeding targets on acid soils [1], which are commonly found in tropical and subtropical regions where sorghum is intensively cultivated. As the primary result of Al toxicity is a damaged and stunted root system, water and nutrient acquisition are thereby compromised, leading to significant yield losses [2].

Because acid soils are widespread in the world [3], Al toxicity represents a major constraint for crop production worldwide, particularly in areas where food security still poses a significant challenge to human populations. For example, Al toxicity and phosphorus deficiency are major constraints for sorghum production in West Africa [4,5]. In addition, yield reduction caused by drought stress, which is common in the region, worsens as roots intoxicated by Al are incapable of penetrating the deep, highly acidic soil layers to acquire water [6]. A potential threat to food

security arises as sorghum and pearl millet are the main staple food crops in the West African Savannah zones [7]. In view of today's challenge of feeding nine billion people in the near future, including the most economically disadvantaged [8], modern molecular strategies are needed [9].

Organic acids such as malate and citrate are released by Al-activated organic acid transporters located in the plasma membrane of cells in the root apex. Once in the rhizosphere, they form stable complexes with Al, thereby conferring Al tolerance [10]. Major Al tolerance genes belonging to the aluminum-activated malate transporter (ALMT) and multidrug and toxic compound efflux (MATE) families were first cloned in wheat (*TaALMT1*, [11]), sorghum (*SbMATE*, [12]) and barley (*HvAACT1*, [13]), and subsequently have been found to control Al tolerance in a number of other plant species (reviewed in [14]).

In sorghum, the major Al tolerance locus, *Alt<sub>SB</sub>*, was mapped to the end of sorghum chromosome 3 and explains 80% of the phenotypic variation in a mapping population derived from the

Al-tolerant line, SC283, and the Al-sensitive line, BR007 [15]. An allelic series at the *All<sub>SB</sub>* locus was found to underlie highly contrasting phenotypes in sorghum. Evidence in support of other distinct Al tolerance genes was also found [16]. Subsequently, high resolution mapping localized *All<sub>SB</sub>* to a 24.6 kb region within which *SbMATE* (GenBank accession EF611342), a gene encoding an aluminum-activated citrate transporter belonging to the MATE family, was found to underlie the Al tolerance locus [12]. *SbMATE* is expressed in the roots of a tolerant near-isogenic line (NIL) in an aluminum-inducible fashion, with highest expression localized to the first centimeter of the root. The *SbMATE* coding region was completely monomorphic between the parental alleles. In conjunction with a significant positive correlation between Al tolerance and both citrate exudation and *SbMATE* expression, this suggests that polymorphisms in regulatory regions underlie the allelic effects at the *All<sub>SB</sub>* locus by modulating *SbMATE* expression. Polymorphisms in the *All<sub>SB</sub>* region included a variable Tourist-like miniature inverted repeat transposable element (MITE) insertion in the promoter region and SNPs and indels located in the second *SbMATE* intron and within two amplicons in the *SbMATE* 3' region [12]. The size of the MITE insertion across different sorghum lines is positively correlated with Al tolerance, suggesting that this insertion harbors *cis*-acting elements that enhance *SbMATE* expression in Al tolerant genotypes. *Cis*-acting elements altering Al tolerance gene expression have been reported for *TaALMT1* in wheat [17], *HvAACT1* in barley [18] and for the *TaMATE1B* gene in wheat [19].

There is growing evidence that regulatory factors modulate the expression of Al tolerance genes. For example, in Arabidopsis, expression of both *AtALMT1* and *AtMATE* has been shown to be regulated by a C2H2-type zinc finger transcription factor, STOP1, which is also associated with tolerance to low pH [20,21]. A homolog of STOP1, ART1, regulates the expression of a suite of genes related to Al tolerance in rice, including *STAR1* and *STAR2* [22], *Nrat1* [23], *OxALSI* (an ABC transporter involved in rice Al tolerance, [24]) and the MATE family member, *OxFRDL4* [25].

Using different donors, *All<sub>SB</sub>* was introgressed into a genetic background belonging to an Al sensitive line [26]. A varying reduction in both Al tolerance and *SbMATE* expression from parents to the derived NILs was observed and the NILs differed for *SbMATE* expression, suggesting that *SbMATE* expression is regulated at multiple levels. That is, these findings suggest that although *cis* effects are dominant in controlling *SbMATE* expression, the loss of functional *trans*-acting factors may lead to potentially strong genetic background effects depending on the donor allele, reducing both *SbMATE* expression and Al tolerance.

Population structure, which is common in sorghum [27,28,29], must be controlled in order to avoid false positives in association genetics research. We have previously reported that Al tolerance is a rare trait in sorghum and is not randomly distributed across the species diversity continuum [30]. That is, Caniato and colleagues found that 80% of the sorghum accessions were Al sensitive, 14% were identified as intermediately tolerant and only 6% were highly Al tolerant. Excluding breeding derivatives, only 5% of the entire panel was found to be highly tolerant to Al. Al tolerance is more prevalent in guinea and to lesser extent caudatum subpopulations, suggesting that causal mutations might have arisen in West Africa, after the guinea race differentiated from the primordial bicolor types.

In the present study, the panel described in [30] was used for association mapping, focusing on the 24.6 kb region where the *All<sub>SB</sub>* locus is located on sorghum chromosome 3. We observed a fast rate of LD decay in the region. Intra-locus recombination events were found to explain much of the haplotypic diversity

observed for *All<sub>SB</sub>* and were instrumental in narrowing down the location of causal variants, which are likely located in a 6 kb region encompassing the MITE insertion and intronic polymorphisms. A haplotype network based on polymorphisms associated with Al tolerance suggests a single, recent origin of Al tolerance conferred by *All<sub>SB</sub>* in guinea sorghums from West Africa. The implication of these findings as a basis for allele mining strategies to identify Al tolerant accessions is discussed.

## Results

### Linkage Disequilibrium in the *All<sub>SB</sub>* Region

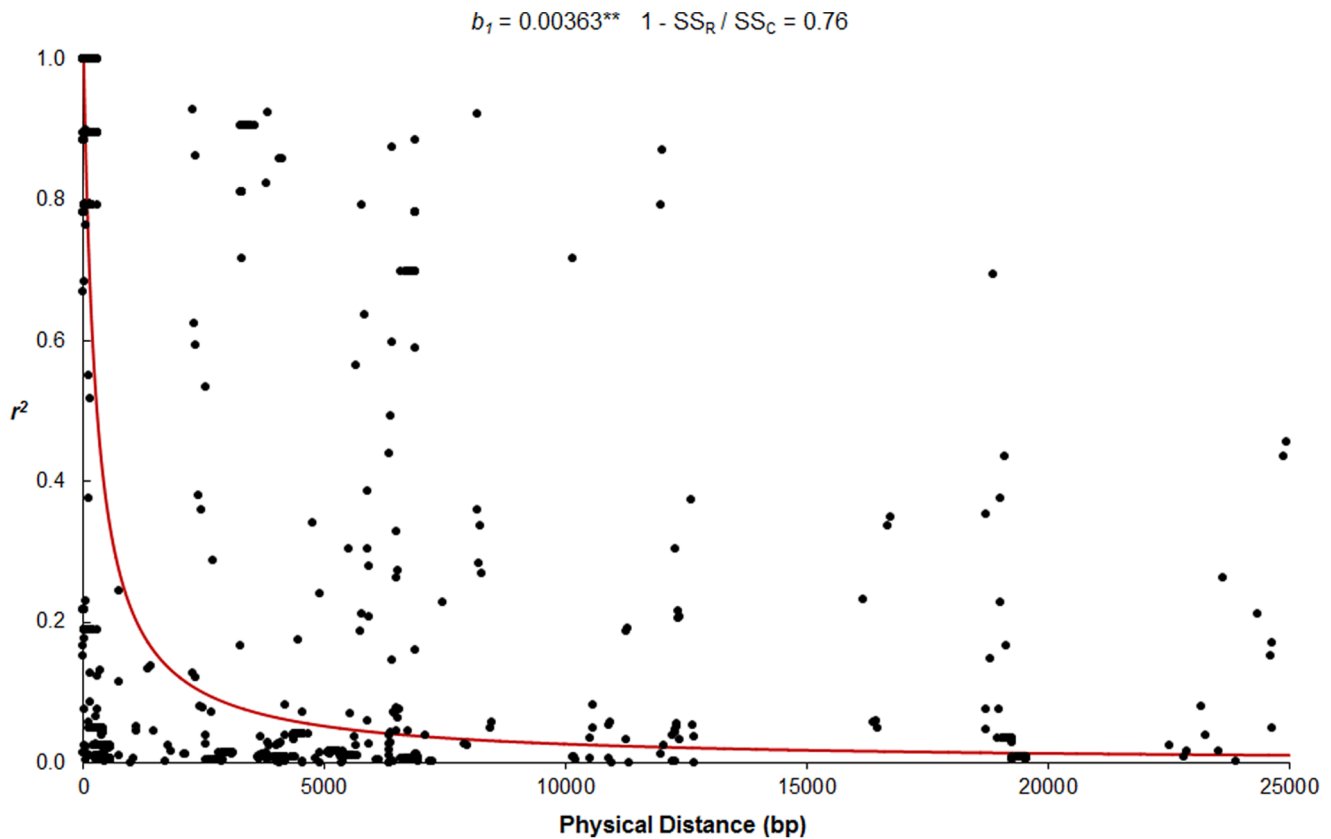
The LD decay model based on drift-recombination equilibrium explained approximately 76% of the variance in  $r^2$  estimates. Figure 1 shows there is a steep decrease in  $r^2$  estimates, which dropped to  $\sim 0.2$  and then close to zero for sites separated by  $\sim 1$  kb and  $\sim 5$  kb, respectively. An analysis of mean  $r^2$  estimates and the respective standard deviations supports rapid LD decay in the *All<sub>SB</sub>* region (Table 1). Although LD decay was less pronounced based on  $D'$ , which is not sensitive to differences in allele frequencies [31], it was also significant using this statistic (regression coefficient,  $b_1 = 0.000028$ ,  $p < 0.0001$ ). Based on  $D'$ , 85% of sites 1 kb apart but only 38% of sites more than 5 kb apart were in significant LD ( $p < 0.05$ ).

### Association Model Fitting and Type I Error Control

For each tested model, the probability distribution under the null hypothesis was obtained by plotting the  $p$ -values resulting from association analysis against the cumulative  $p$ -values (Figure 2). A non-uniform distribution was found for the naïve,  $Q_6$  (six subpopulations) and K models in the cumulative plots, resulting in inflated type I error. Using a variety of approaches we have previously determined that six subpopulations result in a meaningful representation of the genetic diversity patterns in this sorghum association panel [30]. Nonetheless, here we also tested 4, 6, 8 and 12 subpopulations for type I error control. In agreement with our expectations, while the performance of the  $Q_4$  model in error control was reduced relative to  $Q_6$ , no advantage was observed with higher subpopulation numbers (data not shown). Therefore, the most effective control of false positives was achieved with a mixed model that included six subpopulations combined with familial relatedness ( $Q_6+K$ ), which resulted in the fewest false positive associations among all tested models.

### Association Analysis Revealed Polymorphisms Associated with Al Tolerance

Broad-sense heritability estimates for Al tolerance based on RNRG data exceeded 0.9. Using  $p < 0.01$  as a threshold, 14 of the 21 polymorphisms identified within the *All<sub>SB</sub>* region were found to be significantly associated with Al tolerance as represented by *RNRG<sub>5d</sub>* (Figure 3a), with five associated loci located within the second intron of *SbMATE*. As expected, associated loci were in general in LD (Figure 3b), preventing the unambiguous identification of causative polymorphism(s). The SNP locus with the strongest association was 6083 ( $-\log_{10}(p) \cong 15$ ) within the second *SbMATE* intron, explaining approximately 12% of the Al tolerance variation, with an allele substitution effect on Al inhibition of root growth (*RNRG<sub>5d</sub>*) of  $\sim 54\%$  (Figure 3c, Tables S1 and S2). At this locus, the allele associated with Al tolerance, A, was the least frequent allele (minor allele frequency, MAF = 0.10, Table S3 and S4). Loci with  $8 < -\log_{10}(p) < 12$  were 199, the MITE locus (MIV) in the promoter region, with a  $\sim 43\%$  effect on *RNRG<sub>5d</sub>*, marker 6094 in the second *SbMATE* intron and 8364, 8423 and 12487 downstream of *SbMATE*, which each increased



**Figure 1. LD decay in the *Alt<sub>SB</sub>* region.** In red is the prediction obtained by fitting a nonlinear regression model of the squared correlation of allele frequencies ( $r^2$ ) as a function of physical distance between pairs of loci based on the drift-recombination model [40]. The regression coefficient ( $b_1$ ,  $**p < 0.0001$ ) and the fraction of the total variance explained by the nonlinear model ( $1 - SS_R / SS_T$ ) are shown, where  $SS_R$  and  $SS_T$  are the sum of squares of error and total, respectively.  
doi:10.1371/journal.pone.0087438.g001

*RNRG<sub>5d</sub>* by ~60% (Table S2). Loci located on the edges of the *Alt<sub>SB</sub>* region up to the ~12 kb position, such as 161, 199 and 12487, were still in significant LD with the loci located within or close to *SbMATE* and thus are still associated with Al tolerance. Associations became significantly weaker ( $-\log_{10}(p) \cong 2$ ) for polymorphisms located at physical positions relatively distant from *SbMATE*, near the 25 kb position in Figure 3a. Overall, this suggests that causal variants are located between loci 161 and 12487.

### Haplotype Diversity and the Fine Structure of the *Alt<sub>SB</sub>* Locus

Consistent with the steep decrease in LD observed in the *Alt<sub>SB</sub>* region, the four gamete test revealed at least five likely

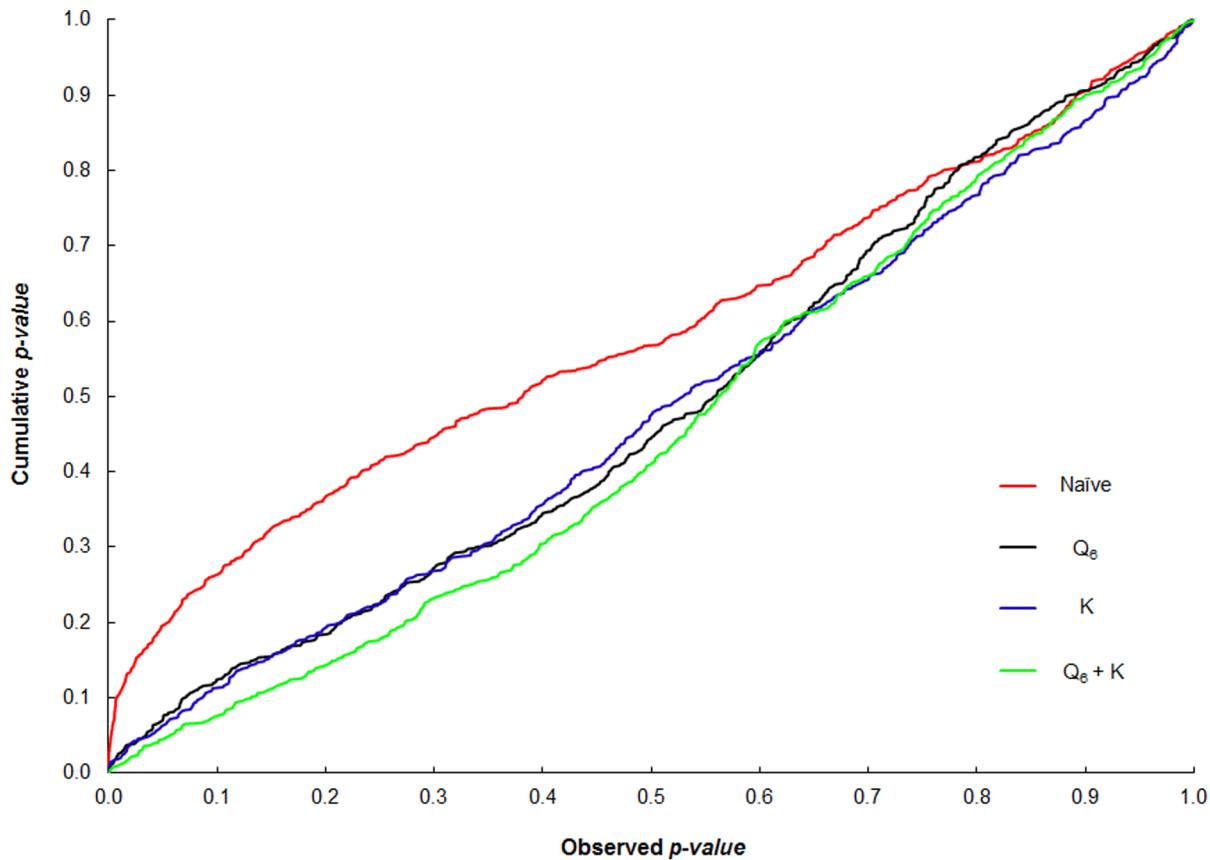
recombination events in the region, with one intragenic recombination event detected in the second intron of *SbMATE* (between loci 6083 and 6097, Figure 4a). Because recombination may create homoplasmy [32], introducing ambiguity into the relationships among *Alt<sub>SB</sub>* haplotypes, a haplotype network was built based on nine loci associated with Al tolerance, comprising eight different haplotypes (Figure 4b). This network shows the mutational relationships among the haplotypes, but it does not necessarily represent the mutational history, which is unknown. The two haplotypes with the highest frequency by far were H1 and H2, which differ only by a T(H2)  $\leftrightarrow$  C(H1) transition at the outer edge of the *Alt<sub>SB</sub>* region (locus 24804), with the C allele being present in all the other haplotypes.

Under neutrality, there is an expected relationship between haplotype frequency and haplotype age [33]. That is, the most common allele is likely to be the oldest with a probability equal to its frequency [34,35]. Therefore, in view of their much higher frequency with respect to the other haplotypes, H1 and H2 appear to be the ancestral types in which Al tolerance mutations arose. Supporting this hypothesis is the high frequency of H1 in the subpopulation Q<sub>5</sub> that includes the bicolor race, which is believed to be the ancestral morphological race from which the other sorghum races originated [36]. The only haplotype composed exclusively of alleles increasing Al tolerance was H5, which carries four alleles not found on any other haplotype. Interestingly, the putative ancestral haplotypes H2/H1 were formed exclusively by alleles associated with Al sensitivity, except for the C allele at the

**Table 1. Mean squared allele-frequency correlations ( $r^2$ ) and respective standard deviation across the *Alt<sub>SB</sub>* region.**

Region (bp)	Number of sites	Mean $r^2$	Standard Deviation
1–500	673	0.73	0.40
501–1024	51	0.02	0.04
1088–5079	289	0.14	0.30
5092–24934	418	0.11	0.22

doi:10.1371/journal.pone.0087438.t001



**Figure 2. Model comparison for type I error control.** Type I error distribution obtained with the naïve,  $Q_e$ , K and  $Q_e + K$  models using 38 SSR loci and phenotypic traits related to Al tolerance. Under the expectation that the randomly distributed SSR loci are not associated with Al tolerance, models that properly control the type I error should show a uniform distribution of  $p$ -values along a diagonal line in the cumulative plot. Loci with MAF >0.1 were used.

doi:10.1371/journal.pone.0087438.g002

24804 locus in H1. However, this locus showed only marginal association with Al tolerance and also had a negligible effect in  $RNRG_{5d}$  (Table S2) indicating that it does not provide significant Al tolerance to the putative ancestral haplotype.

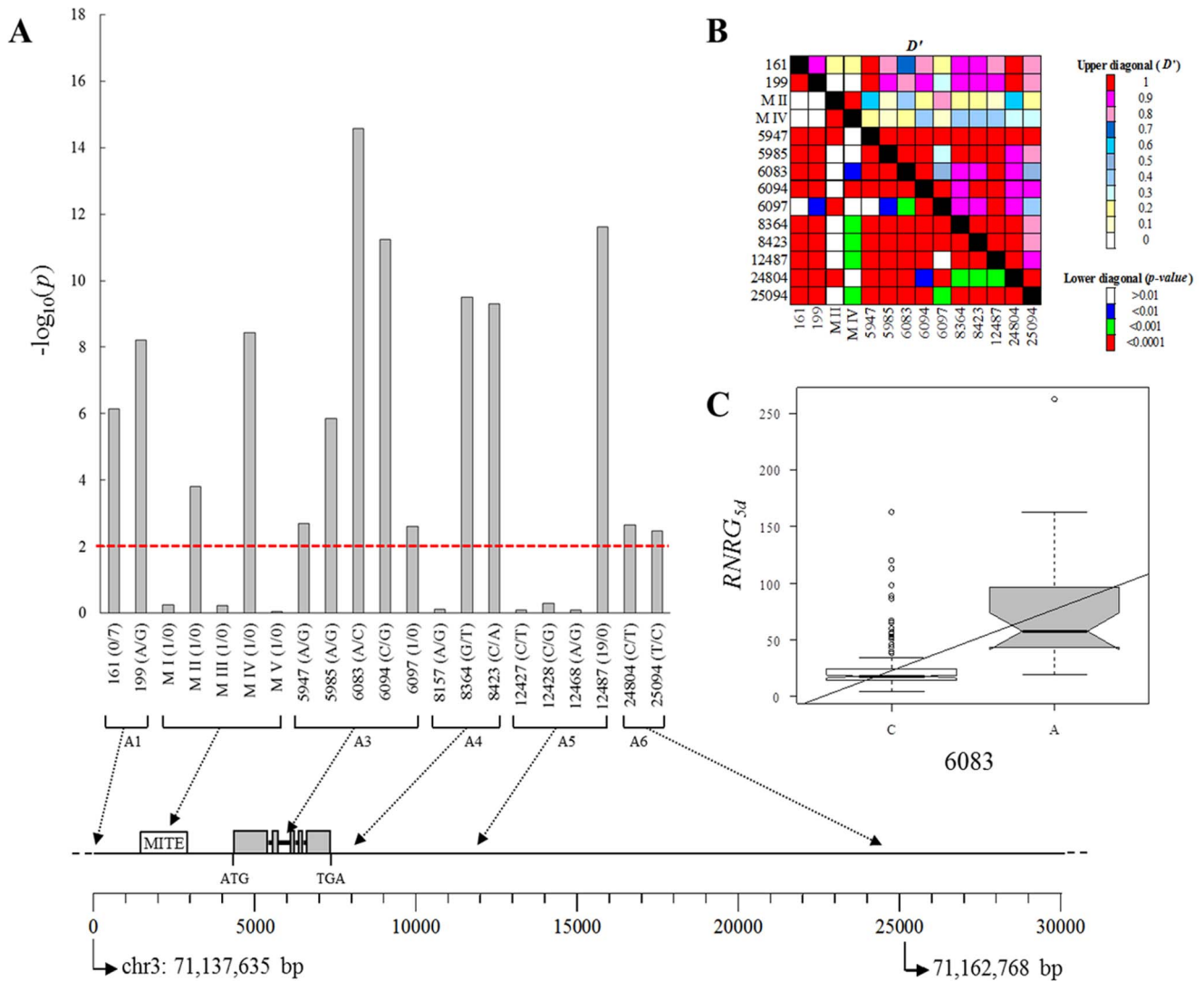
Incompatible splits depicted by loops in the network may represent events such as hybridization, horizontal gene transfer or recombination [37]. The low frequency haplotypes, H4, H6, H7 and H8, which are formed by alleles associated with both Al tolerance and Al sensitivity, are possibly the result of recombination events involving other haplotypes. Based on their higher frequencies and allelic constitution, with alternative alleles at all loci, H1 and H5 may have been involved in recombination events giving rise to H4, H6, H7 and H8. This is supported by the position of the recombination events detected by the four gamete test (see recombination breakpoints in H4, H6, H7 and H8 in Figure 4a).

Jointly considering H1 and H2, the percentage of Al tolerant and intermediate accessions harboring either haplotype was rather low (~11%). In contrast, the vast majority of the accessions carrying the H5 haplotype, ~90%, were either Al tolerant or intermediately tolerant. Significantly differently from the case for H1 and H2, H5 is present exclusively in subpopulations Q1, Q3 and Q6, which are composed primarily of guinea accessions with western (Q1) and southern African and Asian (Q6) origins, as well as breeding derivatives (Q3).

Interestingly, the A  $\leftrightarrow$  C transversion at the 6083 locus appears to correlate well with the occurrence of a high level of Al tolerance in sorghum. Considering the H5 haplotype and the putative recombinant haplotypes in the network loop, high Al tolerance was more frequently found in haplotypes carrying the A allele (eight Al tolerant accessions in H5, H6 and H8) rather than the C allele (only one Al tolerant accession in H3, H4 and H7) at the 6083 locus. Recombination was deemed important for the evolution of the  $\beta$ -globin region in humans [35]. Similarly, intra-locus recombination appears to be an important process underlying the haplotypic diversity for the sorghum Al tolerance locus,  $All_{SB}$ .

### Allele Mining

In the context of allele mining, we then set out to investigate the power of different associated loci to recover Al tolerant and intermediate accessions in the association panel. Table 2 shows that loci combining high  $-\log(p)$  for association with Al tolerance and lower MAFs, between 0.06 and 0.10, were the most efficient ones in recovering Al tolerant accessions in the panel. For example, the loci 6083 and 6094 in the second intron of *SbMATE*, and 8364, 8423 and 12487 downstream of its coding region, recovered between 79 to 86% of Al tolerant and intermediate accessions. Among those, the 6083 locus stands out as it recovers nine out of the 14 highly Al tolerant accessions. Complementarity between associated loci is compromised by the finding that loci



**Figure 3. Association analysis for polymorphisms in the *Alt<sub>SB</sub>* region and Al tolerance.** Association analysis with the  $Q_6+$  K model was performed with  $RNRG_{sd}$ . **(A)** Statistical significance is expressed as  $-\log_{10}(p)$  and the  $p < 0.01$  threshold is represented by the red horizontal line. Polymorphisms are shown along the x-axis and are linked to the schematic below this graph which depicts their physical location in the 24.6 kb region where *SbMATE* (exons shown as gray boxes connected to black lines representing introns) was mapped on chromosome 3 (A1 to A5 depict amplicons harboring polymorphisms, Table S5). The corresponding physical positions in the sorghum genome are shown below the scale and were obtained by sequence similarity analysis (<http://www.phytozome.net>). The alleles at each loci are shown in the x-axis following the locus designation, with indels represented by the number of repeats, except for the MITE insertion, which was coded as described in the Material and Methods session. **(B)** Linkage disequilibrium expressed by pairwise  $D'$  estimates [65] among loci associated with Al tolerance.  $p$ -values obtained with the Fisher exact test are shown. **(C)** Allele substitution effect for the 6083 locus. The slope of the linear regression line indicates an allele substitution effect of 53.9%  $RNRG_{sd}$  ( $p < 2E-16$ ). doi:10.1371/journal.pone.0087438.g003

recovering additional Al tolerant accessions, such as 6097 and 24804 showed rather high MAFs (0.32–0.39), undesirably co-selecting ~80% of Al sensitive accessions.

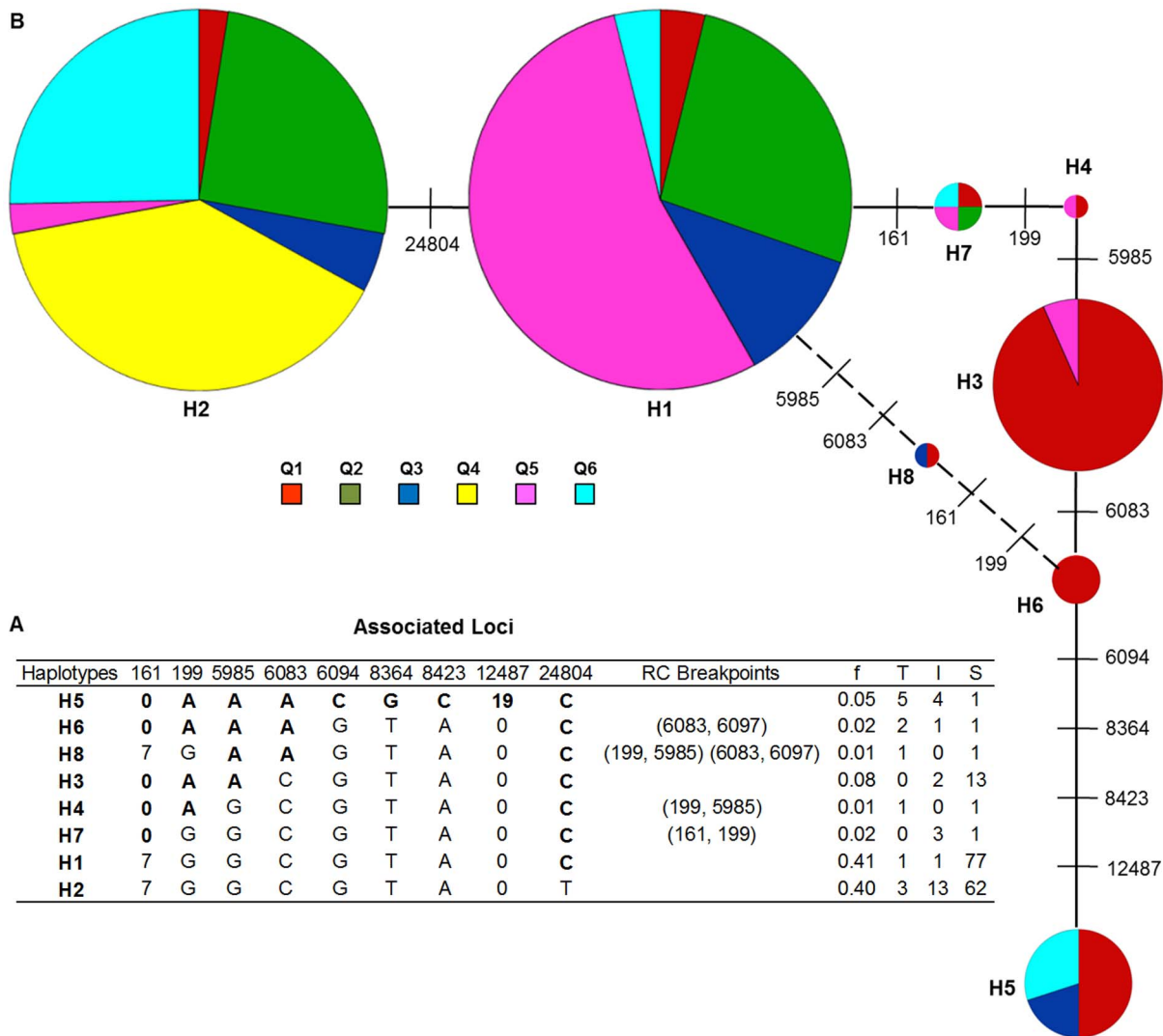
## Discussion

### Fast LD Decay in the *Alt<sub>SB</sub>* Region

Typically, LD decay estimates in sorghum are variable and range from 15–20 kb [38], to 50–100 kb [29] and as high as 150 kb [39]. Based on the drift-recombination model, LD in the *Alt<sub>SB</sub>* region based on  $r^2$  dropped to ~0.2 and close to zero for sites separated by 1 kb and 5 kb, respectively. Strikingly, this rather rapid rate of decay is more comparable to that in the outcrossing

species, maize [40,41,42], than to that in previous reports on the largely self-pollinating sorghum. LD is the complex result of the history of recombination as well as the mutational history [43], which are profoundly affected by demographic factors among others. Interestingly, Lin and colleagues [44] screened 30,000 gametes to fine map the *Shattering1* (*Sh1*) gene to a 17 kb region on sorghum chromosome 1. In comparison, fine mapping of the *Alt<sub>SB</sub>* locus involved the screening of only 4,170 gametes, one seventh of the population size in the *Sh1* study, to resolve *SbMATE* into a 24.6 kb region [12]. Thus, for *Alt<sub>SB</sub>*, local factors appear to be dominant in determining the pattern of LD across the Al tolerance locus. It is possible that the *Alt<sub>SB</sub>* location towards the end of sorghum chromosome 3, which is enriched in gene content [45], is





**Figure 4. Relationship among *Alt<sub>SB</sub>* haplotypes.** (A) Haplotypes with recombination (RC) breakpoints defined by the four-gamete test (two additional possible breakpoints between loci 6097/8364 and 12487/25094 were also detected). Alleles associated with Al tolerance (*RNRG<sub>5d</sub>*) are in bold (at the 161 and 12487 loci, 7/0 and 19/0 depict a 7 bp and a 19 bp indel). Haplotype frequency (f) and the number of Al tolerant (T, *RNRG<sub>5d</sub>* >80%) intermediate (I, 30% < *RNRG<sub>5d</sub>* < 80%) and Al sensitive (S, *RNRG<sub>5d</sub>* < 30%) accessions within each haplotype are shown. (B) *Alt<sub>SB</sub>* network based on nine sites associated with Al tolerance. The circle areas are proportional to haplotype frequencies except for H1 and H2, whose areas were reduced by 2.5-fold due to their much higher frequencies. Colored areas are proportional to the number of accessions within each of the six subpopulations defined in [30]: Q1 (guinea accessions from western Africa and guinea margaritifera accessions), Q2 (caudatum accessions from Africa and group of transplanted caudatum and durra accessions from Lake Chad region), Q3 (lines from the Embrapa collection and US), Q4 (kafir accessions from southern Africa), Q5 (durra accessions from central eastern Africa and from Asia; bicolor and caudatum accessions from Asia), Q6 (guinea accessions from southern Africa and Asia).  
doi:10.1371/journal.pone.0087438.g004

associated with local factors enhancing recombination. In maize, MacMullen and colleagues [46] observed the absence of loci with genome-wide effects on recombination, suggesting the importance of numerous but localized regions affecting recombination, which could be structural chromosome- or family-specific variants.

#### Low Frequency Variants within the *Alt<sub>SB</sub>* Locus are Highly Associated with Al Tolerance

In the present study, we either sequenced or genotyped all polymorphisms previously detected between the parents of the mapping population used to positionally clone *SbMATE*. Causative polymorphisms are thus expected to be included among loci displaying statistically significant associations with Al tolerance.

Association mapping is limited when the trait analyzed is correlated with population structure [47] as is the case of Al tolerance both in sorghum [30] and rice [48]. However, because the proportion of the phenotypic variance explained by population structure alone in our sorghum association panel was only 16% compared to 57% in rice, we expect that a substantial fraction of the phenotypic variance would still be available for capture by Quantitative Trait Nucleotides (QTNs), even in the presence of population membership cofactors in our mixed model. Indeed, 14 loci associated with Al tolerance were found in the *Alt<sub>SB</sub>* region, with 6083 showing the strongest association signal. Because multiple QTNs under LD may control Al tolerance conferred by the *Alt<sub>SB</sub>* locus, it is conceivable that 6083 reflects an indirect

**Table 2.** Recovery of Al tolerant and intermediate accessions by loci associated with Al tolerance.

Locus	Alleles	MAF	n	T	I	S	T+I	S	Al tolerant accessions																
									IS14351	IS21519	IS10801	SC566	SC283	SC175	CMS225	IS26554	IS23142	SDX	IS29691	IS25077	IS26457	Brandes			
161	<b>0/7</b>	0.18	44	8	13	23	48	52	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
199	<b>A/G</b>	0.17	41	8	11	22	46	54	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
MI1	<b>1/0</b>	0.40	103	3	15	85	17	83	+												+				-
MI4	<b>1/0</b>	0.11	25	5	11	9	64	36	+												+				-
5947	<b>G/A</b>	0.47	118	4	21	93	21	79	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5985	<b>A/G</b>	0.19	47	9	12	26	45	55	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
6083	<b>A/C</b>	0.10	24	9	10	5	79	21	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
6094	<b>C/G</b>	0.07	16	6	7	3	81	19	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
6097	<b>0/1</b>	0.32	80	7	8	65	19	81	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
8364	<b>G/T</b>	0.06	14	5	6	3	79	21	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
8423	<b>C/A</b>	0.06	16	6	7	3	81	19	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
12487	<b>19/0</b>	0.06	14	5	7	2	86	14	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
24804	<b>T/C</b>	0.39	152	10	19	123	19	81	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
25094	<b>T/C</b>	0.14	34	4	11	19	44	56	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Phenotypic	-	-	254	17	37	200	21	79																	

In bold are the least frequent alleles, which are linked in coupling with Al tolerance except for loci with borderline association probabilities and/or small effects on *RNRG<sub>sd</sub>* (MI1, 5947 and 24804). For each associated locus are shown the minor allele frequencies (MAF), the total number of accessions (n), the number of Al tolerant (T, *RNRG<sub>sd</sub>* > 80%), intermediate (I, 30% < *RNRG<sub>sd</sub>* < 80%) and sensitive (S, *RNRG<sub>sd</sub>* < 30%) accessions. The proportions based on phenotypic selection (phenotypic) are shown at the bottom of the table. Given their genotypes, the Al tolerant accessions in the association panel marked with '+' possess the Al tolerant *Alt<sub>58</sub>* allele. The Al tolerant lines CMS226, CMS227 and 9929034, which are breeding derivatives from SC283 (Table S1 in [16]), were excluded from this analysis.  
doi:10.1371/journal.pone.0087438.t002

association with one or more causal polymorphism(s) in LD with this locus.

The proportions of Al tolerant and intermediate accessions in the association panel were 6 and 14%, respectively [30]. We observed that loci with extreme association probabilities ( $\log_{10}(p) > 9$ ) showed a rather narrow MAF range, between 0.06 and 0.11, with 6083 showing MAF = 0.10 (Table S3) and  $-\log_{10}(p) \cong 15$  for association with Al tolerance. Higher frequency common variants under LD with presumptive causal variants for the *Sh1* gene in sorghum led to synthetic association signals stronger than those of the causal variants [44]. Here, polymorphisms falling within the frequency interval for Al tolerant and intermediate accessions produced the strongest association signals. For example, although MAFs for loci in the more distant regions with respect to *SbMATE*, such as 161 (MAF = 0.18, Table S3) and 199 (0.17) in the 5' region and 24804 (0.39) and 25094 (0.14) in the 3' region were higher than that for 6083, the respective association signals were much lower. Because the Al tolerance range for intermediate accessions was rather broad ( $30\% < RNRG_{5d} < 80\%$ ), it is likely that in addition to the Al tolerant accessions, only the fraction of the intermediate accessions that are most Al tolerant within that intermediate category contribute significantly to the association signals. As previously reported, the power to detect a QTL is expected to be the highest when the associated marker allele has a similar frequency to that of the QTL [49,50].

### Causal Variants are Likely Localized in a 6 kb Region Encompassing a Variable MITE Insertion and Intronic Polymorphisms

The detection of extensive intra-locus recombination events (Figure 4b) provides support for fast dissipation of LD across the *Alt<sub>SB</sub>* region as revealed by fitting the drift-recombination model for LD decay. These recombinants may help to narrow down further the location of the causal variants affecting *SbMATE* expression.

Based on the proportion of Al tolerant and intermediate accessions within the Al tolerant haplotype, H5, and in the putatively recombinant haplotypes, the causative polymorphism(s) enhancing *SbMATE* expression and thus Al tolerance are likely located at or upstream of the 6083 locus within the second intron of *SbMATE*, including the MITE insertion in the promoter region. This leads to a reduction of the physical interval where causative mutations lie from previous 24.6 kb to only ~6 kb. The location of the causal variants is supported by the observation that 8 of the 13 Al tolerant accessions in our haplotype network carried either the Al tolerant haplotype, H5, or recombinant haplotypes which retained either part (H8) or the whole (H6) region from H5 delimited by the loci, 161 and 6083, with all these accessions showing the H5 allele (A) at the 6083 locus. Among the accessions harboring H5 are the Al tolerant parent of the mapping population used to positionally clone *SbMATE*, SC283, and IS14351, which is the most Al tolerant accession in the panel [30], with both accessions showing high levels of *SbMATE* expression. Another highly Al tolerant line that shows high *SbMATE* expression is SC566 [30], which shows the recombinant haplotype, H8.

Of the remaining five Al tolerant accessions with the C allele at the 6083 locus, one accession carried the H4 haplotype, whereas one and three accessions carried the putative ancestral haplotypes, H1 and H2, respectively. In such a diverse panel, non-allelic heterogeneity in the form of different Al tolerance genes controlling distinct physiological mechanisms of Al tolerance may occur, as has been previously reported in rice for genes regulating panicle length [51]. In fact, we have previously reported

both on allelic and non-allelic heterogeneity for Al tolerance based on the sorghum *Alt<sub>SB</sub>* locus [16, 26, 30]. Our previous studies with NILs indicated extensive allelic heterogeneity in sorghum presumably arising from variations in *cis*-factors within *Alt<sub>SB</sub>* [26]. Therefore, the co-occurrence of weak *Alt<sub>SB</sub>* alleles showing low *SbMATE* expression in conjunction with other Al tolerance genes in the genetic background may account for the presence of Al tolerant accessions with the H1 and H2 haplotypes, which are formed by alleles linked in repulsion with Al tolerance. In fact, we have previously reported on non-allelic heterogeneity for Al tolerance in two accessions carrying the H2 haplotype, 5DX [16] and IS29691 [30]. Our data indicate that 5DX possesses a rather weak *Alt<sub>SB</sub>* allele and that the introgression of different Al tolerant gene(s) into the 'DX' genetic background is likely to account for its high levels of Al tolerance [16]. The *SbMATE* expression level in SC175, the remaining accession with the H2 haplotype, and in IS25077 that harbors H4, was approximately 4-fold smaller than that in the highly Al tolerant line, SC283 (see Figure 3 in [30] for data on IS25077 and Figure 3 in [26] for SC175), suggesting a similar situation in these lines to that observed in 5DX. Finally, although showing high *SbMATE* expression, marker-trait association with markers tightly linked to *Alt<sub>SB</sub>* in a backcross family derived from the single Al tolerant accession harboring the H1 haplotype, IS23142, was not significant (Figure 3 and Table S3, [30]). This could be due to the recessive mode of gene action for Al tolerance observed in this accession and/or the presence of other Al tolerance genes in its background. Therefore, our data suggest the presence of distinct, non *Alt<sub>SB</sub>*-based physiological mechanisms of Al tolerance in the putative ancestral haplotypes and in H4. Prevalence of accessions showing these mechanisms is expected in H1 and H2 due to the high frequency of these haplotypes in the association panel.

On the other hand, although the H5 haplotype was almost entirely composed of either Al tolerant or intermediate accessions, one Al-sensitive accession was found carrying this otherwise totally Al tolerant haplotype. One possible explanation is the significant occurrence of genetic background effects on Al tolerance controlled by *SbMATE* [26]. Accordingly, these background effects are the result of accessory loci acting in *trans*, without which *SbMATE* expression and Al tolerance can be dramatically reduced.

### A Possible Unified Origin of Al Tolerance Conferred by *Alt<sub>SB</sub>* in West African Guinea Types

The landrace accessions used in this study were chosen to provide representation of cultivated landrace sorghums from the whole world, with sampling based on race classification, latitude of origin, response to day length, and form of cultivation [52]. A core reference set including a large proportion of these accessions was found to capture ~80% of the SSR alleles detected in a larger, 3367-member collection, indicating good representation of the global genetic diversity in sorghum [53]. This landrace collection has been recently characterized with over ~265,000 SNP markers to investigate genomic patterns of diversification in sorghum [39].

The haplotypic diversity for *Alt<sub>SB</sub>* based on loci associated with Al tolerance provides insights into the origin of the mutations conferring Al tolerance in sorghum. The haplotype H5 was present exclusively in subpopulations Q1, Q3 and Q6, which are composed primarily of guinea accessions with western (Q1) and southern African and Asian (Q6) origins, as well as breeding derivatives (Q3). With a model selection procedure based on the Bayesian Information Criterion [54], we have previously shown that these three subpopulations are the most important ones in explaining the variation in Al tolerance across the association



panel [30]. In addition, Q1 membership was the most prevalent in H5 accessions and was also frequent in haplotypes in the network loop, where recombination events involving H5 may have taken place.

In conjunction with the much higher frequency of the putative ancestral haplotypes H1/H2 compared to H5, the data presented here support a more recent origin of Al tolerance in the primary domestication center of the guinea race, in West Africa [36]. Given that differentiated haplotypes closely related to H5 but predominantly found in Q6 were not observed in the dataset, Al tolerant haplotypes in the secondary domestication center in South/East Africa are likely to have been transported from West Africa during the guinea radiation from the West to the southern domestication center of the guinea race [55,56]. Therefore, the scenario described here suggests a single geographic and racial origin of Al tolerance mutations within *Alt<sub>SB</sub>* in primordial guinea domesticates in West Africa, with subsequent limited interracial spread of Al tolerance. This is consistent with our previous studies indicating non-random distribution of Al tolerance in the sorghum genetic diversity continuum [30].

### Analysis of the Power of Single SNPs versus *Alt<sub>SB</sub>* Haplotypes for Allele Mining in Sorghum

One important issue is whether single SNPs or haplotypes would be the most effective type of markers for allele mining based on *Alt<sub>SB</sub>*. Considering the associated loci with allele frequency closely matching that of Al tolerance (6083, 6094, 8364, 8423 and 12487), there would be no clear advantage in using haplotypes over the single SNPs for allele mining, as little or no complementation can be exploited in recovering Al tolerant accessions (note that the 6083 locus alone recovers most of the Al tolerant accessions in the panel). This is in line with the observation in barley that when the causal SNP is one of the genotyped markers, the power of single SNPs is superior to that of haplotypes [49]. The consequence of using loci with unmatched allele frequencies, such as 6097 which is only 14 bp from 6083, is the recovery of a large number of Al sensitive accessions, probably due to extensive recombination with the causal variant(s). However, allele mining based solely on 6083 would also recover ~20% of the Al sensitive accessions in the panel, which is likely the result of genetic background effects reducing *SbMATE* expression as previously reported [26].

It is possible that multiple causal variants contribute to Al tolerance conferred by *Alt<sub>SB</sub>*. For example, the strong association signal observed for 6083 may reflect a role for the second *SbMATE* intron in enhancing gene expression. Introns are known to increase mRNA levels by acting as transcriptional enhancers, among other mechanisms [57,58,59]. In addition, here we show that the MITE insertion region in the *SbMATE* promoter is associated with Al tolerance. This result suggests a role for the transposon insertion in enhancing *SbMATE* expression, as has been observed in wheat Al tolerance involving *TaMATE1B* [19]. We are currently working to identify both the *cis* factors enhancing *SbMATE* expression and the *trans*-acting factors involved in genetic background effects.

The associated loci reported in this study are now available for high throughput germplasm screening through the Integrated Breeding Platform from the Generation Challenge Programme (<https://www.integratedbreeding.net/>). Using this Platform with the data presented here should allow us to build a molecular pipeline to increase yield stability for sorghum cultivated on acidic, Al toxic soils.

## Materials and Methods

### Plant Material

A set of 209 accessions from the landrace collection described in [52] and an additional 45 inbred lines formed the association panel that was used in this study. The landrace collection is representative of the genetic diversity present in cultivated sorghum [52], whereas the inbred lines are frequently used in breeding programs in the US and Brazil, including highly Al tolerant donors [16].

### Assessment of Al Tolerance in Nutrient Solution

Al tolerance data collection was undertaken previously [30]. A total of 254 sorghum accessions were evaluated in nutrient solution containing {0} or {27}  $\mu\text{M}$   $\text{Al}^{3+}$ . Values inside brackets indicate  $\text{Al}^{3+}$  activity, which was estimated with the speciation software program, GEOCHEM-EZ [60,61].

Seeds of each genotype were germinated for four days and seedlings were transferred to containers with nutrient solution lacking Al at pH 4.0. After 24 h of acclimation, the initial length of each seedling's primary root growing in control solution (*ilc*) was measured. The solution was then replaced with nutrient solution of identical composition but containing either no Al or {27}  $\mu\text{M}$   $\text{Al}^{3+}$  supplied as  $\text{AlK}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ . Final root lengths under Al treatment (*flAl*) or in control solution (*flc*) were obtained after three and five days of exposure to Al. For each inbred line, relative percent values of net root growth inhibition after three (*RNRG<sub>3d</sub>*) and five (*RNRG<sub>5d</sub>*) days of Al exposure were estimated by dividing the net root growth under Al treatment (*flAl-ilc*) by the net root growth without Al (*flc-ilc*). We adopted here the same Al tolerance classification described previously [30] for the sorghum accessions: Al sensitive (*RNRG<sub>5d</sub>* <30%), intermediately tolerant (30% < *RNRG<sub>5d</sub>* <80%, designated intermediate) and Al tolerant (*RNRG<sub>5d</sub>* >80%).

### Candidate Quantitative Trait Nucleotides (QTNs)

We previously sequenced the entire 24.6 kb *Alt<sub>SB</sub>* region in the Al tolerant and sensitive parents, SC283 and BR007, and identified candidate QTNs (depicted in Figure 1d in [12]). For the association analysis conducted here, we sequenced six amplicons spanning the QTN regions including a T $\leftrightarrow$ A transversion in the first exon of *SbMATE* [26]. The genomic structure for the MITE insertion was previously obtained in four sorghum genotypes (Figure 3e and Supplementary Figure 3 in [12]). In the present study, the MITE insertion was genotyped in the association panel as five biallelic (presence/absence) loci. The four previously detected MITE alleles were designated MI (456 bp), MII (1,184 bp), MIII (1,514 bp), MIV (1,912). An additional 2,280 bp allele detected only in the association panel was designated MV.

### PCR and DNA Sequencing

Leaf tissues from three plants of each accession were used for DNA isolation according to [62]. Amplifications were carried out in a reaction volume of 20  $\mu\text{L}$  that contained 30 ng of genomic DNA, 10X polymerase chain reaction buffer containing 0.5 mM dNTP, 4 mM  $\text{MgCl}_2$ , 10 pmol of each primer, 5% of dimethyl sulfoxide (DMSO) and 1 U of Taq DNA polymerase (Phoentria, Belo Horizonte, MG), following the amplification conditions described in Table S5 that also contains the sequences for the primers used in this study. PCR products were treated with 0.6 U of shrimp alkaline phosphatase (SAP, SB Corporation, Cleveland, OH) and 1.5 U of EXO I (USB Corporation, Cleveland, OH) in a reaction volume of 10  $\mu\text{L}$ , which contained 6  $\mu\text{L}$  of PCR

reactions, 0.4  $\mu$ L 10X SAP buffer (USB Corporation, Cleveland, OH). Digestion proceeded with incubation at 37°C for 30 min followed by 80°C for 10 min for enzyme inactivation. Sequencing reactions were carried out in a 10  $\mu$ L reaction volume which contained 5  $\mu$ L of digested PCR products, 2  $\mu$ L of Big Dye V3.1 (Applied Biosystems, Foster City, CA), 2  $\mu$ L of 5X buffer (Applied Biosystems, Foster City, CA) and 5 pmol of each primer. Sequencing reactions proceeded at 96°C for 4 min, 30 cycles at 96°C for 10 sec, 50°C for 5 sec and 60°C for 4 min. Sequencing reactions were analyzed on a ABI3100 sequencer (Applied Biosystems, Foster City, CA). Sequences were aligned and manually edited using the software SEQMAN (DNASTar, Madison, WI). MITE polymorphisms were scored on 1% (w/v) agarose gels.

### Analysis of Population Structure

Population structure (Q) was previously estimated based on 38 SSR loci which are evenly distributed across the sorghum genome, from a sorghum SSR kit ([53], [http://sat.cirad.fr/sat/sorghum\\_SSR\\_kit/](http://sat.cirad.fr/sat/sorghum_SSR_kit/)) developed within the Generation Challenge Programme (GCP, <http://www.generationcp.org/>). The complete description of the methods used for these analyses in addition to the SSR primer sequences and amplification conditions can be found in [30]. Briefly, the Bayesian cluster analysis as implemented in the software STRUCTURE [63,64] was used to estimate the number of subpopulations based on the SSR data set. The admixture model with correlated allele frequencies was adopted, with a burn-in period of 100,000 and a run length of 1,000,000, with five independent replications for each k (number of subpopulations).

### Linkage Disequilibrium (LD)

LD between polymorphisms with minor allele frequency exceeding 0.05 was estimated using the standardized disequilibrium coefficient ( $D'$ , [65]) and squared allele-frequency correlations ( $r^2$ , [66]) using the TASSEL software program (<http://www.maizegenetics.net/bioinformatics/tasselindex.htm>). Loci in significant LD based on  $D'$  were defined with the Fisher's exact test ( $p < 0.05$ ). The decay of LD with physical distance (bp) was estimated using nonlinear regression (PROC NLIN, SAS® software, SAS Institute Inc., Cary, NC, USA) based on the drift-recombination model described in [40]. Accordingly, the expected value of  $r^2$  under drift-recombination equilibrium is  $E(r^2) = 1/(1+C)$ , where  $C = 4Nc$  ( $N$  is effective population size and  $c$  is the recombination fraction between loci, [67]).

### Association Models

We tested three models to control for false positive associations (type I error) using the 38 randomly distributed SSR loci and various Al tolerance data:  $RNRG_{3db}$ ,  $RNRG_{5db}$ , Visual Root Damage (VRD), Induction of Root Growth (IRG) and Principal Components (PC) 1 and PC2, which are described in [30]). This marker density does not provide genome saturation within the low LD context in the association panel [29] and consequently, the chances of association with the phenotypic traits can be considered negligible. Thus, association analysis with these markers provides a null distribution to test the efficiency at which different models control for false positive associations. The naïve model, which does

not account for familial relatedness or kinship, is  $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{e}$ ; the Q model, which accounts for population structure, is  $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{Q}\mathbf{v} + \mathbf{e}$ ; the K model, which accounts for familial relatedness or kinship, is expressed as  $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ . In these models,  $\mathbf{y}$  is a vector of phenotypic observations,  $\boldsymbol{\alpha}$  is a vector of fixed effects related to SNP effects (QTNs),  $\mathbf{e}$  is a vector of residual effects,  $\mathbf{v}$  is a vector of fixed effects related to population structure and  $\mathbf{u}$  is a vector of polygene background random effects related to familial relatedness.  $\mathbf{A}$  and  $\mathbf{Z}$  are the incidence matrices of 0s and 1s, relating  $\boldsymbol{\alpha}$  and  $\mathbf{u}$ , respectively, to  $\mathbf{y}$ .  $\mathbf{Q}$  is the population membership assignment matrix obtained from the software, STRUCTURE relating  $\mathbf{y}$  to  $\mathbf{v}$ . Finally,  $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{Q}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ , the unified Q+K mixed model that jointly accounts for population structure and familial relatedness [68] was fitted to the data. The variances of the random effects are expressed as  $\mathbf{Var}(\mathbf{u}) = 2\mathbf{K}\mathbf{Vg}$  and  $\mathbf{Var}(\mathbf{e}) = \mathbf{R}\mathbf{Vr}$ , where  $\mathbf{K}$  is a 254×254 matrix based on the proportion of shared alleles values [69], obtained with the PowerMarker software [70],  $\mathbf{R}$  is a 254×254 matrix with the off-diagonal elements being zero and the diagonal elements being the reciprocal of the number of observations for which each phenotypic data point was obtained, and  $\mathbf{Vg}$  and  $\mathbf{Vr}$  are the genetic and residual variance, respectively.

### Haplotype Diversity and Network

A simplified haplotype network was constructed by maximum parsimony, using nine biallelic sites that were associated with Al tolerance, and omitting singleton haplotypes. The four-gamete test [71] was applied to identify possible recombination events.

### Supporting Information

#### Table S1 Association statistics for loci in the *Alt<sub>SB</sub>* region.

(DOC)

#### Table S2 Allele substitution effects on $RNRG_{5db}$ (%) for loci associated with Al tolerance.

(DOC)

#### Table S3 Minor allele frequencies for loci in the *Alt<sub>SB</sub>* region.

(DOC)

#### Table S4 Complete genotypic and phenotypic dataset.

(XLSX)

#### Table S5 Primer sequences and amplification conditions.

(DOC)

### Acknowledgments

We would like to thank Dr. Jean-François Rami from CIRAD – France for supplying the landrace panel used in this study.

### Author Contributions

Conceived and designed the experiments: JVM. Performed the experiments: FFC. Analyzed the data: FFC JVM MH CTG ZZ. Contributed reagents/materials/analysis tools: RES. Wrote the paper: FFC JVM MH CTG RES ZZ LVK.

### References

- Garvin D, Carver B (2003) Role of the genotype in tolerance to acidity and aluminum toxicity. In Rengel Z, ed. Handbook of Soil Acidity. New York: Marcel Dekker. 387–406.
- Kochian LV, Hoekenga OA, Piñeros MA (2004) How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. Annual Review of Plant Biology 55: 459–493.

3. von Uexküll HR and Mutert E (1995) Global extent, development and economic impact of acid soils. *Plant and Soil* 171: 1–15.
4. Doumbia MD, Hossner LR, Onken AB (1993) Variable sorghum growth in acid soils of subhumid West Africa. *Arid Land Research and Management* 7: 335–346.
5. Doumbia MD, Hossner LR, Onken AB (1998) Sorghum growth in acid soils of West Africa: variations in soil chemical properties. *Arid Land Research and Management* 12: 179–190.
6. Foy CD, Duncan RR, Waskom RM, Miller DR (1993) Tolerance of sorghum genotypes to an acid, aluminum toxic tatum subsoil. *Journal of Plant Nutrition* 16: 97–127.
7. Weltzien E, Rattunde HFW, Clerget B, Siart S, Toure A, et al. (2006) Sorghum diversity and adaptation to drought in West Africa. In: Jarvis D, Mar I, Sears L, editors. *Enhancing the use of crop genetic diversity to manage abiotic stress in agricultural production systems*. Rome: International Plant Genetic Resources Institute. 31–38.
8. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, et al. (2010) Food security: the challenge of feeding 9 billion people. *Science* 327: 812–818.
9. Fedoroff NV, Battisti DS, Beachy RN, Cooper PJM, Fischhoff DA, et al. (2010) Radically rethinking agriculture for the 21st century. *Science* 327: 833–834.
10. Delhaize E, Gruber BD, Ryan PR (2007) The roles of organic anion permeases in aluminium resistance and mineral nutrition. *FEBS Letters* 581: 2255–2262.
11. Sasaki T, Yamamoto Y, Ezaki B, Katsuhara M, Ahn SJ, et al. (2004) A wheat gene encoding an aluminum-activated malate transporter. *The Plant Journal* 37: 645–653.
12. Magalhaes JV, Liu J, Guimarães CT, Lana UGP, Alves VMC, et al. (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nature Genetics* 39: 1156–1161.
13. Furukawa J, Yamaji N, Wang H, Mitani N, Murata Y, et al. (2007) An aluminum-activated citrate transporter in barley. *Plant Cell Physiology* 48: 1081–1091.
14. Liu J, Luo X, Shaff J, Liang C, Jia X, et al. (2012) A promoter-swap strategy between the *AtALMT* and *AtMATE* genes increased Arabidopsis aluminum resistance and improved carbon-use efficiency for aluminum resistance. *The Plant Journal* 71: 327–337.
15. Magalhaes JV, Garvin DF, Wang Y, Sorrells ME, Klein PE, et al. (2004) Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the Poaceae. *Genetics* 167: 1905–1914.
16. Caniato FF, Guimarães CT, Schaffert RE, Alves VMC, Kochian LV, et al. (2007) Genetic diversity for aluminum tolerance in sorghum. *Theoretical and Applied Genetics* 114: 863–876.
17. Ryan PR, Raman H, Gupta S, Sasaki T, Yamamoto Y, et al. (2010) The multiple origins of aluminium resistance in hexaploid wheat include *Aegilops tauschii* and more recent *cis* mutations to *TaALMT1*. *The Plant Journal* 64: 446–455.
18. Fujii M, Yokosho K, Yamaji N, Saisho D, Yamane M, et al. (2012) Acquisition of aluminium tolerance by modification of a single gene in barley. *Nature Communications* 3: 713.
19. Tovkach A, Ryan PR, Richardson AE, Lewis DC, Rathjen TM, et al. (2013) Transposon-mediated alteration of *TaMATE1B* expression in wheat confers constitutive citrate efflux from root apices. *Plant Physiology* 161: 880–892.
20. Iuchi S, Koyama H, Iuchi A, Kobayashi Y, Kitabayashi S, et al. (2007) Zinc finger protein STOP1 is critical for proton tolerance in Arabidopsis and coregulates a key gene in aluminum tolerance. *Proceedings of the National Academy of Sciences* 104: 9900–9905.
21. Liu J, Magalhaes JV, Shaff J, Kochian LV (2009) Aluminum-activated citrate and malate transporters from the MATE and ALMT families function independently to confer Arabidopsis aluminum tolerance. *The Plant Journal* 57: 389–399.
22. Yamaji N, Huang CF, Nagao S, Yano M, Sato Y, et al. (2009) A zinc finger transcription factor ART1 regulates multiple genes implicated in aluminum tolerance in rice. *The Plant Cell* 21: 3339–3349.
23. Xia J, Yamaji N, Kasai T, Ma JF (2010) Plasma membrane-localized transporter for aluminum in rice. *Proceedings of the National Academy of Sciences* 107: 18381–18385.
24. Huang C-F, Yamaji N, Chen Z, Ma JF (2012) A tonoplast-localized half-size ABC transporter is required for internal detoxification of aluminum in rice. *The Plant Journal* 69: 857–867.
25. Yokosho K, Yamaji N, Ma JF (2011) An Al-inducible MATE gene is involved in external detoxification of Al in rice. *The Plant Journal* 68: 1061–1069.
26. Melo JO, Lana UGP, Piñeros MA, Alves VMC, Guimarães CT, et al. (2013) Incomplete transfer of accessory loci influencing *SbMATE* expression underlies genetic background effects for aluminum tolerance in sorghum. *The Plant Journal* 1: 276–288.
27. Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, et al. (2008) Community resources and strategies for association mapping in sorghum. *Crop Science* 48: 30–40.
28. de Alencar Figueiredo LF, Calatayud C, Dupuits C, Billot C, Rami J-F, et al. (2008) Phylogeographic evidence of crop neodiversity in sorghum. *Genetics* 179: 997–1008.
29. Bouchet S, Pot D, Deu M, Rami J-F, Billot C, et al. (2012) Genetic structure, linkage disequilibrium and signature of selection in sorghum: lessons from physically anchored DArT markers. *PLoS One* 7: e33470.
30. Caniato FF, Guimarães CT, Hamblin M, Billot C, Rami J-F, et al. (2011) The relationship between population structure and aluminum tolerance in cultivated sorghum. *PLoS One* 6: e20830.
31. Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331–341.
32. Smouse PE (1998) To tree or not to tree. *Molecular Ecology* 7: 399–412.
33. Posada D, Crandall K (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* 16: 37–45.
34. Waterson GA, Guess HA (1977) Is the most frequent allele the oldest? *Theoretical Population Biology* 160: 141–160.
35. Long JC, Chakravarti A, Boehm CD, Antonarakis S, Kazazian HH (1990) Phylogeny of human  $\beta$ -globin haplotypes and its implications for recent human evolution. *American Journal of Physical Anthropology* 81: 113–130.
36. de Wet MJ (1978) Systematics and evolution of sorghum sect. Sorghum (Gramineae). *American Journal of Botany* 65: 477–484.
37. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.
38. Hamblin MT, Fernandez MGS, Casa AM, Mitchell SE, Paterson AH, et al. (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. *Genetics* 171: 1247–1256.
39. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences* 110: 453–458.
40. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences* 98: 11479–11484.
41. Yu J, Buckler E (2006) Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* 17: 155–160.
42. Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, et al. (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 4: e8451.
43. Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* 18: 83–90.
44. Lin Z, Li X, Shannon LM, Yeh C-T, Wang ML, et al. (2012) Parallel domestication of the *Shattering1* genes in cereals. *Nature Genetics* 44: 720–724.
45. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457: 551–556.
46. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. *Science* 325: 737–740.
47. Larsson SJ, Lipka AE, Buckler ES (2013) Lessons from *Dwarf6* on the strengths and weaknesses of structured association mapping. *PLoS Genetics* 9: e1003246.
48. Famoso AN, Zhao K, Clark RT, Tung C-W, Wright MH, et al. (2011) Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genetics* 7: e1002221.
49. Lorenz AJ, Hamblin MT, Jannink J-L (2010) Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One* 5: e14079.
50. Hamblin MT, Jannink J-L (2011) Factors affecting the power of haplotype markers in association studies. *The Plant Genome* 4: 145–153.
51. Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, et al. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature communications* 2: 467.
52. Deu M, Rattunde F, Chantreaux J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. *Genome* 49: 168–180.
53. Billot C, Ramu P, Bouchet S, Chantreaux J, Deu M, et al. (2013) Massive sorghum collection genotyped with SSR markers to enhance use of global genetic resources. *PLoS One* 8: e59714.
54. Schwarz B (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
55. de Wet MJ, Harlan JR, Kurmarohita B (1972) Origin and evolution of guinea sorghums. *East African Agricultural and Forestry Journal* 38: 114–119.
56. Harlan JR, de Wet MJ (1972). A simplified classification of cultivated sorghum. *Crop Science* 12: 172–176.
57. Rose AB, Beliakoff JA (2000) Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiology* 122: 535–542.
58. Nott A, Meislin SH, Moore MJ (2003) A quantitative analysis of intron effects on mammalian gene expression. *RNA* 9: 607–617.
59. Rose AB (2008) Intron-mediated regulation of gene expression. *Current Topics in Microbiology and Immunology* 326: 277–290.
60. 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, editors. *Chemical Equilibrium and Reaction Models*. Madison: Soil Science Society of America. 253–269.
61. Shaff JE, Schultz BA, Craft EJ, Clark RT, Kochian LV (2009) GEOCHEM-EZ: a chemical speciation program with greater power and flexibility. *Plant and Soil* 330: 207–214.
62. Saghai-Marroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromo-

- somal location, and population dynamics. *Proceedings of the National Academy of Sciences* 81: 8014–8018.
63. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
  64. 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.
  65. Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49–67.
  66. Hill WG, Robertson A (1966) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38: 226–231.
  67. Sved JA (1971) Linkage disequilibrium of chromosome segments. *Theoretical Population Biology* 2: 125–141.
  68. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38: 203–208.
  69. Jin L, Chakraborty R (1994) Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. *Molecular Biology and Evolution* 11: 120–127.
  70. Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128–2129.
  71. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164.