High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array

Wubishet A. Bekele1, Silke Wieckhorst2, Wolfgang Friedt1 and Rod J. Snowdon1,*

1Department of Plant Breeding, Justus Liebig University, Giessen, Germany
2KWS Saat AG, Einbeck, Germany

Summary
With its small, diploid and completely sequenced genome, sorghum (Sorghum bicolor L. Moench) is highly amenable to genomics-based breeding approaches. Here, we describe the development and testing of a robust single-nucleotide polymorphism (SNP) array platform that enables polymorphism screening for genome-wide and trait-linked polymorphisms in genetically diverse S. bicolor populations. Whole-genome sequences with 6× to 12× coverage from five genetically diverse S. bicolor genotypes, including three sweet sorghums and two grain sorghums, were aligned to the sorghum reference genome. From over 1 million high-quality SNPs, we selected 2124 Infinium Type II SNPs that were informative in all six source genomes, gave an optimal Assay Design Tool (ADT) score, had allele frequencies of 50% in the six genotypes and were evenly spaced throughout the S. bicolor genome. Furthermore, by phenotype-based pool sequencing, we selected an additional 876 SNPs with a phenotypic association to early-stage chilling tolerance, a key trait for European sorghum breeding. The 3000 attempted bead types were used to populate half of a dual-species Illumina iSelect SNP array. The array was tested using 564 Sorghum spp. genotypes, including offspring from four unrelated recombinant inbred line (RIL) and F2 populations and a genetic diversity collection. A high call rate of over 80% enabled validation of 2620 robust and polymorphic sorghum SNPs, underlining the efficiency of the array development scheme for whole-genome SNP selection and screening, with diverse applications including genetic mapping, genome-wide association studies and genomic selection.

Keywords: Illumina, Infinium, single-nucleotide polymorphism, SNP chip.

Introduction
Sorghum is widely grown as a staple cereal crop, particularly in Africa and parts of Asia but also for various uses in other parts of the world. In Australia and the US Southern Plains, grain sorghum represents a drought-tolerant alternative to maize production for livestock feeding, while in Europe, China and North America, interest is also growing rapidly in the use of sweet and/or grain sorghum forms as a potentially drought-tolerant and nutrient-efficient alternative to maize for bio-energy production (Rooney et al., 2007). Breeding sorghum for temperate regions necessitates the combination of chilling tolerance from appropriate germplasm resources with photoperiod adaptation and other appropriate agronomic characters (e.g. high dry-matter biomass or grain yields), whereas in subsistid and semi-arid farming systems, grain yield and seed quality are the most vital traits. In comparison with more established crops like maize, wheat and barley, breeding of sorghum for bio-energy and livestock feeding is a relatively young enterprise, and an enormous genetic potential for improvement in the crop has yet to be tapped by breeders. The ability to intercross cultivated sorghum races with related subspecies for expansion of genetic diversity and improvement in key traits is a unique aspect of this crop (Washburn et al., 2011).

The small diploid genome of sorghum, the availability of a completed reference genome sequence (Paterson et al., 2009) and the consequent ability to develop cost-effective, high-throughput tools for whole-genome screening make sorghum a strikingly amenable crop for the application of genomics-based breeding methods. In particular, the dramatically falling costs of genome-wide screening for single-nucleotide polymorphisms (SNPs), using high-density SNP array technologies (see Batley and Edwards, 2007) or genotyping-by-sequencing (Chia and Ware, 2011; Davey et al., 2011; Elshire et al., 2011; Morris et al., 2013) on next-generation sequencing platforms (Metzker, 2010), has opened the way for genomic selection (Jannink et al., 2010) or predictive breeding strategies (Riedelsheimer et al., 2012). Such techniques have the potential to considerably accelerate selection gain and improve the effectiveness of breeding.

Nelson et al. (2011) used reduced-representation sequencing of restriction-site-associated DNA (RAD; see Baird et al., 2008) to discover genome-wide SNPs in a panel of 8 genetically diverse grain sorghum genotypes. By whole-genome resequencing of two sweet sorghum accessions and one grain sorghum, Zheng et al. (2011) not only detected over 1 million genomic SNPs, but also demonstrated that over 1500 genes differentiate between sweet and grain sorghum. This gene diversity spans important processes of high relevance for breeding, for example sugar and starch metabolism, lignin biosynthesis and stress responses. This demonstrates that a broad genetic basis of genome-wide sequence variation is necessary to capture SNP variation of general relevance for genetic diversity and breeding towards different end-use scenarios. In the present study, we supplemented the sequence data from Zheng et al. (2011) by resequencing one further sweet sorghum and one further grain sorghum genotype, both of geographically and genetically

---

doi: 10.1111/pbi.12106
divergent origin. Furthermore, we generated two phenotypic pools of 30 recombinant inbred lines (RILs) each, segregating strongly for early-stage chilling tolerance, from a cross between the latter two genotypes. These 60 RILs were skim-sequenced with a moderate genome coverage to identify trait-linked SNPs of relevance to European energy sorghum breeding. Using six whole-genome sequences (including the *S. bicolor* reference sequence from the grain sorghum BTx623), we were able to detect an extremely high number of high-quality genome-wide SNPs with high allele frequencies in genetically diverse *S. bicolor* germplasm, encompassing both grain and sweet sorghum types. The SNP array we developed using this data (Figure 1) was used to validate 2620 robust and polymorphic SNPs in a panel of 564 genetically diverse *S. bicolor* accessions, including four unrelated RIL and F$_2$-$3$ mapping populations and a *Sorghum* spp. genetic diversity collection.

**Results and discussion**

**Paired-end short sequence mapping**

The technological advancement of paired-end sequencing makes mapping of short DNA sequence reads onto the sorghum genome extremely precise and efficient despite the large proportion of repeat units (Paterson *et al.*, 2009; Zheng *et al.*, 2011). In contrast, Nelson *et al.* (2011) applied single-end Illumina sequencing for SNP detection sequencing in eight diverse sorghum lines. Our resequencing strategy mapped more than 90% (Table 1) of the total reads onto the reference genome, comparable to the frequency of successfully mapped reads reported by Zheng *et al.* (2011); with single-end reads from restriction-site-associated DNA (RAD) sequences only 30% of the reads could be correctly mapped by Nelson *et al.* (2011).

**Resequencing in sweet and grain sorghums**

The huge quantity of SNPs we detected in SS79, M71, Keller, E-tian and Ji-2731 underlines the high level of polymorphism present in sorghum. In the study of Zheng *et al.* (2011), Ji-2731 was reported to show the largest SNP diversity compared with BTx623; however, we revealed that many of the identified polymorphisms were heterozygote state SNPs. The patterns of SNP distribution we observed across the genome reflected other reports on sorghum and other plants (Morris *et al.*, 2013; Paterson *et al.*, 2009; Zheng *et al.*, 2011). Low SNP densities were seen in centromeric regions, with high numbers of SNPs towards the chromosome ends (Figure 2). As expected, the sweet sorghum genotypes SS79 and Keller exhibited greater SNP diversity compared with the grain sorghum reference sequence than was observed in the other grain sorghum genotypes (Table 2). Interestingly, significant genome-scale difference between the three sweet sorghums and the three grain sorghum types (including BTx623) was observed in a large chromosome block on chromosome Sb10 (25-45 Mb) (Figure 2). On the other hand, we observed a region of approximately 25 Mbp on chromosome Sb02 with low polymorphism in all five accessions compared with BTx623, while a 35 Mbp region on chromosome Sb04 was strongly divergent in the two grain sorghum genotypes we analysed compared with the three sweet sorghums and the BTx623 grain sorghum reference. More detailed study of such regions can potentially give interesting insights into sorghum domestication and adaptation to diverse climatic and agricultural systems. For example, quantitative trait loci (QTL) influencing sorghum maturity or photoperiod sensitivity, and with a pleiotropic effect on sugar, biomass and grain yield, have been identified in biparental sweet x grain sorghum populations within the corresponding region of chromosome Sb04 (Murray *et al.*, 2008; Shiringani *et al.*, 2010). In a nested association mapping study using more than 1000 individuals from 24 families, Mace and Jordan (2013) identified 40 small-effect QTL showing synteny to flowering-time QTL in maize. Their results revealed hotspots for flowering-time QTL on chromosomes Sb03 and Sb04. Some of the candidate genes they reported lie within the interesting region of differentiation that we identified. The most prominent of these is Sb04g008320 (*SbFT*; 9.47Mbp), a

![Figure 1](image-url) **Figure 1** Outline of the single-nucleotide polymorphism (SNP) detection, filtering and array development scheme. Single-nucleotide polymorphism calling was performed with CLC Genomics Workbench and SHOREmap, respectively. Details of the phenotypic bulk SNP identification are described in Experimental procedures.
were also detected by CLC, whereby only 70% of SHOREmap showed that more than 90% of all SNPs called by SHOREmap and SHOREmap in the two parental genotypes SS79 and M71 genome. It will help to better understand the genetic basis of sweet and metabolism (Kang and Turano, 2003). The presence of deletions and copy-number variations in vital genes like the transport, light stimulus response and regulation of C:N ratio homology to Arabidopsis GLR related gene and has an ionotropic glutamate receptor (GLR) gene involved in calcium ion transport, light stimulus response and regulation of C:N ratio and metabolism (Kang and Turano, 2003). The presence of significant and widespread structural variation in sweet sorghum genomes compared with the grain sorghum reference sequence underlines the need for a dedicated, de novo sweet sorghum reference genome for future genome research. This will help to better understand the genetic basis of sweet and grain sorghum differentiation. It will also provide direct access to gene sequences and regulatory regions that are specific to grain sorghum differentiation. It will also provide direct access to gene sequences and regulatory regions that are specific to grain sorghum differentiation.

### Table 1: Comparison of SNPs detected using CLC genomics between the parental lines SS79 and M71, and between two phenotypic pools of 30 SS79 × M71 RILs each, with high tolerance and high sensitivity to early-stage chilling stress, respectively

<table>
<thead>
<tr>
<th>Parental lines in comparison with BTx623</th>
<th>SS79</th>
<th>M71</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reads after trim</td>
<td>48 712 177</td>
<td>44 167 614</td>
</tr>
<tr>
<td>Total mapped reads</td>
<td>48 512 190</td>
<td>40 586 315</td>
</tr>
<tr>
<td>Fraction of reference covered</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Average coverage excluding zero coverage regions</td>
<td>6.51</td>
<td>5.78</td>
</tr>
<tr>
<td>SNPs in genic regions</td>
<td>133 094</td>
<td>112 185</td>
</tr>
<tr>
<td>SNPs in noncoding regions</td>
<td>1 603 454</td>
<td>778 829</td>
</tr>
<tr>
<td>Total bi-allelic SNPs</td>
<td>1 184 364</td>
<td>636 307</td>
</tr>
<tr>
<td>Bi-allelic SNPs/kb</td>
<td>1.6</td>
<td>0.86</td>
</tr>
<tr>
<td>Deletions</td>
<td>195 165</td>
<td>220 776</td>
</tr>
<tr>
<td>Amplifications</td>
<td>114 772</td>
<td>104 363</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenotype pools in comparison with BTx623</th>
<th>Chilling-tolerant pool</th>
<th>Chilling-sensitive pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mapped reads</td>
<td>170 983 983</td>
<td>189 267 873</td>
</tr>
<tr>
<td>Fraction of reference covered</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Average coverage excluding zero coverage regions</td>
<td>23.16</td>
<td>25.62</td>
</tr>
<tr>
<td>SNPs identified with relaxed settings</td>
<td>2 245 063</td>
<td>3 885 829</td>
</tr>
<tr>
<td>SNPs identified with stringent settings</td>
<td>184 470</td>
<td>1 999 839</td>
</tr>
<tr>
<td>SNPs unique to the phenotypic pool</td>
<td>673</td>
<td>7264</td>
</tr>
<tr>
<td>Unique SNPs in genic regions</td>
<td>84</td>
<td>831</td>
</tr>
</tbody>
</table>

The 673 unique SNPs associated with chilling resistance were found to be dispersed across the whole-genome; however, specific regions of chromosomes 2, 4, 5 and 8 contained more than 25 SNPs/10 Mbp that were unique to the chilling stress phenotype. Interestingly, 12.9% of the SNPs unique to the two phenotypic pools were found to be genic markers, although the SNP selection was not biased towards coding sequences. A total of 6591 or 3.2% were detected that were unique within the susceptible pool. The extremely low number of unique SNPs associated with the phenotypic pool representing RILs with high chilling stress resistance demonstrates the enormous power of bulked-segregant genome-wide sequencing for detection of markers tightly linked to QTL for agronomic traits in crops. A similar bulked segregant analysis was used by Hu et al. (2012) to identify a gene for pod shattering in rapeseed, for example.

The 673 unique SNPs associated with chilling resistance were found to be dispersed across the whole-genome; however, specific regions of chromosomes 2, 4, 5 and 8 contained more than 25 SNPs/10 Mbp that were unique to the chilling stress phenotype. Interestingly, 12.9% of the SNPs unique to the two phenotypic pools were found to be genic markers, although the SNP selection was not biased towards coding sequences. A total of 831 genic SNPs were unique to the chilling-susceptible pool and 84 to the chilling-resistant pool. A GO-term annotation enrichment test revealed enrichment of genes involved in death, growth response to stimuli and antioxidant activity.

### Selection of markers for array development

Flanking 60 nt sequences for 76 574 SNPs were uploaded into the illumina Assay Design Tool for final selection of SNPs to be included on the iSelect BeadChip. From 3000 attempted bead types, a total of 2620 successful beads were generated (Data S1) for SNPs with more or less even distribution across the genome (Figure 3). The average distance between markers was 252.59 kbp, with a minimum distance of 15 bp and a maximum distance of 4688.33 kbp between a pair of SNPs in the centromeric region of chromosome Sb01. The number of markers per chromosome corresponds to the chromosome lengths; for example, the longest sorghum chromosome Sb02 was represented with the highest number of 338 SNPs per chromosome. The successful beads were generated in combination with SNPs from *Allium* species. During the design of the array, the sorghum and *Allium* SNPs were coselected to minimize interspecific cross-hybridization. Interrogation of the arrays and clustering was downstream selection and array development, we only used SNPs that were called by both programmes with quality scores of at least Q40. A total of 163,027 SNPs were selected that had quality scores ≥ Q40 in SS79 and M71 and were also called in all six genotypes with an allele frequency of 50% (Figure 1).
nevertheless performed separately with sorghum and Allium targets, and all Allium SNPs were zeroed during the clustering of the sorghum SNPs.

**Clustering and call rates**

From 3000 attempted bead types, a total of 2620 (87.3%) resulted in successful assays and up to 93.2% of these could be successfully called in the plant materials we screened. This conversion rate of 81.4% corresponds to the expected average design conversion rate of 80% for Illumina’s Infinium assays. The observed call rates ranged from 77.8 to 93.2 in the different populations we screened, giving an average call rate in the 576 samples of 88.96%.

The average ADT score of all attempted sorghum SNP bead types, including failed bead types and SNPs with a call rate of zero, was 0.98. Hence, the ADT score was ruled out as a potential reason for SNP failure at any stage. The reproducibility error rates, calculated using the control sample (SS79) replicated in all plates, were well within the Illumina quality specification (≤ 0.005) defined for validated human SNP Infinium genotyping.

**SNP array characterization and comparison of SNP detection methods**

A total of 760 (29%) of the SNPs in our array panel were derived from the bulked segregant sequencing, while the remaining 1860...
(71%) were genome-wide SNPs. From the latter SNPs, 76% (1415) were found to have a minor allele frequency of more than 1% in the screening panel, whereas only 56.3% (428) of the SNPs derived from the bulked sequencing had a minor allele frequency greater than 1%. The lower polymorphism in the bulk-derived SNPs can be explained by the lower diversity that was considered in selecting these SNPs, where only the polymorphism within SS79 × M71 RIL population in comparison with the reference genome was used rather than six available genomes used for the genome-wide selection of SNPs with a 50% allele frequency. Natural or artificial selection at a locus is known to cause reduced diversity and increased Linkage disequilibrium (LD) (Oraguzie et al., 2007). Because most of the lines screened in this study are part of a breeding programme for Germany, these lines might have been indirectly selected for these genes, and as a result, the markers could not be polymorphic for these loci. One nice example for such reduced diversity due to selection in sorghum is the case of low diversity in sorghum conversion lines (long day insensitive lines) near the maturity locus on chromosome 6 (Morris et al., 2013). Detailed validation of the failed and monomorphic markers using PCR-based methods or other SNP array technologies could give better insight into the complex segregation and inheritance of such SNPs, especially in the case of the SNPs derived from the bulk sequencing analysis.

Despite the reduction in polymorphism in the bulk-derived SNPs, these still showed a good conversion and quality rate. This high accuracy of the SNP detection presumably arose from the parallel validation of the SNPs detected in SS79 and M71 using two alternative SNP calling algorithms (CLC and SHOREmap). The use of genome-wide sequence data for SNP selection also enables an extremely stringent quality score filtering, because the number of SNPs is a virtually nonlimiting factor. Selection of markers developed from diverse lines and resources has been applied with success to develop SNP arrays for numerous species, including cattle, maize, chickpea and tomato (Ganal et al., 2011; Hiermath et al., 2012; Matukumalli et al., 2009; Sim et al., 2012). It is worth noting here that the array design scheme we followed gave a very high rate of successful and polymorphic markers; hence, the same scheme can readily be scaled up for the development of higher density arrays using the same SNP selection list. This opens the possibility for different scenarios, for example selection of SNP panels to target recombination-rich genome regions at higher density for genomic selection or genome-wide association studies.

Array polymorphism and its application for population classification

After automated clustering in the 564 genotypes (Table S1) using the default cluster file, only 127 (4.85%) of the 2620 scorable markers were found to have a call rate of 0 (PIC = 1). Visual screening of the image files revealed almost perfect clustering for all SNPs into the three clusters expected for a simple diploid organism like sorghum. Hence, we made no alterations to the cluster file except for removal of failed SNPs. A total of 1843 (70.34%) SNPs (MAF > 0.01) were polymorphic at the entire

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of markers</th>
<th>Minimum distance (kbp) between markers</th>
<th>Mean distance (kbp) between markers</th>
<th>Maximum distance (kbp) between markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sb01</td>
<td>278</td>
<td>2.70</td>
<td>265.32</td>
<td>4848.33</td>
</tr>
<tr>
<td>Sb02</td>
<td>338</td>
<td>0.10</td>
<td>230.87</td>
<td>2178.96</td>
</tr>
<tr>
<td>Sb03</td>
<td>296</td>
<td>0.65</td>
<td>251.26</td>
<td>4549.23</td>
</tr>
<tr>
<td>Sb04</td>
<td>255</td>
<td>1.28</td>
<td>267.53</td>
<td>4553.42</td>
</tr>
<tr>
<td>Sb05</td>
<td>285</td>
<td>2.55</td>
<td>218.82</td>
<td>2758.09</td>
</tr>
<tr>
<td>Sb06</td>
<td>196</td>
<td>0.02</td>
<td>316.75</td>
<td>3289.38</td>
</tr>
<tr>
<td>Sb07</td>
<td>200</td>
<td>0.84</td>
<td>322.93</td>
<td>4091.32</td>
</tr>
<tr>
<td>Sb08</td>
<td>255</td>
<td>1.87</td>
<td>217.45</td>
<td>2725.55</td>
</tr>
<tr>
<td>Sb09</td>
<td>264</td>
<td>1.87</td>
<td>217.45</td>
<td>3327.44</td>
</tr>
<tr>
<td>Sb10</td>
<td>253</td>
<td>1.87</td>
<td>217.45</td>
<td>2725.55</td>
</tr>
<tr>
<td>Total</td>
<td>2620</td>
<td>0.02</td>
<td>252.59</td>
<td>4848.33</td>
</tr>
</tbody>
</table>

Figure 3 Genome distributions of the 3000 single-nucleotide polymorphisms (SNPs) selected for the array. The x axis indicates 5 Mbp intervals along the ten Sorghum bicolor chromosomes, while the y axis represents the frequency of selected SNPs within each 5 Mbp bin.
screening array level (Table 3). The average PIC value obtained within 1574 polymorphic markers in the diversity panel \((n = 208)\) was 0.20, with a range from 0.0096 to 0.65. From the total of 1683 (64.2%) SNPs, which had a minor allele frequency (MAF) greater than 0.01 within the diversity panel \((n = 208)\), 1574 (85.4%) markers had a minimum call rate of 80%. Detailed SNP array information on call rate and allele frequencies of SNP markers at a whole genotype set \((n = 564)\), polymorphism in the mapping population \((n = 92)\) and the diversity panels \((n = 208)\) and its subpopulations \((n = 154\) and \(n = 54)\) are listed on the Supplementary Table S1.

Figure 4 shows a neighbour joining dendrogram constructed using 1843 SNPs with MAF > 0.01. The dendrogram clearly distinguishes the four mapping populations and the diversity panel. The latter showed the expected high diversity, described previously by Fiedler et al. (2012) using Diversity Array Technology (DArT) markers, while the clustering of the segregating populations reflects the genetic relationships among the parental lines with regard to the variation spanned by the diversity panel.

**Population structure and linkage disequilibrium**

A total of 1574 SNPs (with MAF > 0.01 and a minimum call rate of 80%) were used for population structure analysis on the diversity panel \((n = 200)\) and in the parental lines from the mapping populations \((n = 8)\). The \(\Delta K\) derived from the structure analysis reveals the best cluster at \(K = 2\), dividing the panel into two genetically different subpopulations with 154 and 54 genotypes, respectively (Figure S1). This supports the results of a previous analysis of 194 accessions from the diversity panel using DArT markers, which also revealed two subpopulations (Fiedler et al., 2012). A net nucleotide distance of 0.1240 was revealed between the clusters, with expected heterozygosity of 0.08 within subpopulation/cluster 1 and 0.30 within subpopulation/cluster 2. The parental lines from the four mapping populations all grouped into the second subpopulation contributing to the higher diversity observed within the cluster.

Patterns of linkage disequilibrium were studied in the same set of genotypes \((n = 208)\), using the same 1574 markers on the entire panel as well as within the two subpopulations separately using the corresponding markers with MAF > 0.01 and minimum call rate of 80% (Table 3). A mean \(r^2\) value of 0.052 was calculated for the entire diversity set, while the smaller subpopulation 2 (\(n = 54\)) showed an intermediate mean \(r^2\) value of 0.047 and the larger subpopulation had a mean \(r^2\) value of only 0.034. However, the number of polymorphic loci between the 54 lines in the smaller subpopulation was higher than in both the larger subpopulation and the whole population.

The average LD decay \((r^2 = 0.05)\) reflects the high level of recombination in sorghum, which was already reported by several authors (e.g. Bouchet et al., 2012; Morris et al., 2013). Nevertheless, a closer look at LD decay on a subpopulation level showed that the LD decay in subpopulation 2 (with only 54 lines) decays slightly faster than the larger subpopulation 1 and the entire population (Figure 5). Using the formula of Breseghello and Sorrells (2006), the critical \(r^2\) value calculated for the whole population on chromosome Sb01 was 1.976. The LD starts to decay after just a few bp; however, the distance at which the nonlinear regression line intercepts the critical value is around 400 kb. This represents quite a large extent of LD compared with the recently published results of Morris et al. (2013), which estimated the LD in sorghum to decay at 75–150 kb depending on the genomic region. In that study, however, many more SNPs were analysed in a collection of over 950 worldwide sorghum lines representing several races and agro-climatic conditions. In comparison, most of the lines in our study (154/208) were in one subpopulation with lower level of diversity. Hence, the higher LD results from the lower diversity, as for example found by Lu et al. (2011) in temperate maize compared with tropical maize germplasm.

As reported previously in many plant and animal genomes, not only were differences in LD observed between subpopulations and chromosomes, but also across the length of chromosomes. This is explained by the higher recombination frequencies at the distal ends of the sorghum chromosomes (Mace and Jordan, 2011). The highest levels of LD were found within the heterochromatic region surrounding the centromere. Recombination suppression rates of up to 33% were reported by Kim et al. (2005) in sorghum heterochromatin. For high-resolution genome-wide association studies (GWAS), the high level of LD decay in gene-rich regions of sorghum diversity panels calls for the use of panels with hundreds of thousands of markers. Morris et al. (2013) achieved this marker density using a genotyping-by-sequencing (GBS) approach, whereas whole-genome resequencing as performed in this study is able to potentially reveal all genome-wide SNPs. Genomic skim sequencing may become a viable alternative for GBS as costs for next-generation sequencing continue to fall.

On the other hand, scalable high-density SNP arrays continue to be the method of choice for many applications where large numbers of specific loci need to be repeatedly assayed in large plant populations. This is particularly the case for research organizations and companies without the bioinformatics capacity to deal with GBS data. When genotyping for genomic selection during crop breeding, an increase in the quantity and density of SNPs with GBS does not necessarily increase the accuracy of selection or prediction models. It is often more important to increase the training population size, meaning that low-cost genotyping of large sample numbers for a fixed set of loci may be the priority. Recently, for example, a demonstration of the application of GBS for genomic selection in wheat showed that a

**Table 3** Summary of polymorphism rates in the tested populations among the SNPs on the 3k SNP Infinium array

<table>
<thead>
<tr>
<th>Minor allele frequency (MAF)</th>
<th>Minimum SNP call rate (%)</th>
<th>Entire screening panel ((n = 564))</th>
<th>Complete diversity set ((n = 208))</th>
<th>Diversity set subpop. (1 (n = 154))</th>
<th>Diversity set subpop. (2 (n = 54))</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.01</td>
<td>–</td>
<td>1843 (71%)</td>
<td>1683 (64%)</td>
<td>1683 (64%)</td>
<td>1838 (70%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1738 (66%)</td>
<td>1574 (60%)</td>
<td>846 (32%)</td>
<td>1699 (65%)</td>
</tr>
<tr>
<td>&gt;0.05</td>
<td>–</td>
<td>1795 (69%)</td>
<td>968 (37%)</td>
<td>903 (34%)</td>
<td>1510 (58%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1694 (65%)</td>
<td>882 (34%)</td>
<td>529 (20%)</td>
<td>1640 (63%)</td>
</tr>
</tbody>
</table>

© 2013 Society for Experimental Biology, Association of Applied Biologists and John Wiley & Sons Ltd, Plant Biotechnology Journal, 11, 1112–1125
reduction from 35,000 to 2000 GBS markers did not significantly reduce the prediction accuracy. On the other hand, the 2000 GBS markers did predict performance more accurately than the same number of DArT markers; this was believed to result from a clustering of many DArT markers to certain regions of the genome, whereas GBS markers are more randomly dispersed (Poland et al., 2012). In this regard, our SNP array is also expected to be advantageous for genomic selection and predictive breeding applications, because the selection strategy resulted in an even distribution of SNPs throughout the entire genome (Figure 3).

The diversity panel tested in the present study was previously genotyped with DArT markers by Fiedler et al. (2012), who were only able to use 171 polymorphic loci for population structure and association mapping studies. The genome-wide SNPs we used increased the coverage with high-quality markers more than ninefold. Individual breeding programmes for many crop plants have a relatively low effective population size ($N_e$) due to long-term selection and limited access to adapted lines (e.g. with regard to day length dependency or low temperature tolerance in sorghum). As a result, many genomic selection studies reach a plateau of accuracy at only a few thousand markers. This means that the hundreds of thousands of SNPs that can potentially be obtained from GBS methods (e.g. Heffner et al., 2011; Jannink et al., 2010; Morris et al., 2013; Poland et al., 2012; Zhao et al., 2012) may not necessarily increase prediction potential and may in fact unnecessarily increase the computational complexity of prediction models.
A high-density sorghum genetic map

The parental lines for the SS79 × M71 RIL population segregated for 1198 SNP markers. A very large proportion of these SNPs (1163, or 97.1%) were successfully placed onto a linkage map derived from 92 F$_2$ RILs. Regression ordering of the linkage groups resulted in an average marker density of around 1 marker per cM, demonstrating the usefulness of the array for development of dense linkage maps with sequence annotations to the sorghum genome sequence (Table 4 and Figure 6). The physical and linkage map orders of the markers corresponded in most cases (Figure 7); whereby, the conservation of the physical and genetic marker orders was greatly improved by linkage analysis using the maximum likelihood mapping (MLM) function of Joinmap 4 rather than regression mapping. Maximum likelihood mapping is reported to be extremely sensitive to missing data, however (Cheema and Dicks, 2009; Jansen et al., 2001), causing inflated genetic distance estimates. This explains the exaggerated total length of the map we calculated by MLM, which covered more than 2161.14 cM compared with only 1068 for the map calculated by regression. On the other hand, the improved marker order achieved by the MLM function could be particularly clearly shown for chromosomes Sb01 and Sb03 (Figure 7). Similar physical and genetic mapping colinearity were reported in maize and tomato (Ganal et al., 2011; Sim et al., 2012). Only 17 markers (1.46%) were assigned to another chromosome than their projected positions. Six of these SNPs gave additional blast hits (e-10) to the same chromosome, suggesting possible duplications, while the remainder had multiple blast hits on several chromosomes.

Chromosomes Sb04 and Sb07 were each fragmented into three linkage groups; however, the marker orders within the linkage groups were colinear with the presumed physical positions. Similar fragmentation of linkage groups was reported previously during high-density genetic mapping (Hiermath et al., 2012); however, this is generally not expected to impede accurate QTL mapping. Analyses of allele frequencies in the SM-RIL mapping population indicate a region of distorted allele frequencies on the long arm of chromosome Sb04, which probably led to the difficulties in linkage mapping in this region. The mapping parents also show low levels of polymorphism in a large block on chromosome Sb07, which was reported by Morris et al. (2013) to have low heterozygosity in sorghum.

Finally, our use of large numbers of more or less equidistant markers meant that we were able to accurately map local relationships between genetic and physical map distance. This is another important consideration when considering strategies for map-based cloning of target genes. As in maize, tomato and chickpea maps generated using high-density arrays (Ganal et al., 2011; Hiermath et al., 2012), many heterochromatic regions showed lower recombination frequencies compared with the euchromatic regions.

### Conclusions

Availability of the reference sorghum genome sequence has paved the way for low-cost resequencing and identification of genome-wide SNPs that can potentially enhance genetic analysis and the application of molecular markers in sorghum genomics and breeding. Comparison of different SNP detection strategies revealed the feasibility of detecting high-quality, highly polymorphic SNPs even with low coverage sequencing. Furthermore, we demonstrated the enormous power of phenotypic pool sequencing for detection of trait-associated or QTL-linked SNP panels. Alignment to the genome sequence of a high-density genetic map, containing many equidistant SNPs, provided basic knowledge that will be useful to identify and characterize candidate regions for map-based cloning, using case-by-case experimental designs that reflect local LD levels. Analysis of 564 genetically diverse sorghum accessions including a diversity panel and different mapping populations revealed the usefulness of our SNP array for forward genetic analysis and genomic selection strategies in sorghum breeding populations.

### Table 4 Summary of the SM-RIL Sorghum bicolor linkage map generated with the regression marker order function. A total of 35 additional polymorphic markers remained unlinked

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of markers</th>
<th>Distance (cM) regression marker order</th>
<th>Marker interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sb01</td>
<td>122</td>
<td>77.973</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb02</td>
<td>134</td>
<td>114.094</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb03, a</td>
<td>4</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb03, b</td>
<td>116</td>
<td>87.824</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb04, a</td>
<td>22</td>
<td>36.632</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb04, b</td>
<td>74</td>
<td>72.705</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb04, c</td>
<td>15</td>
<td>4.307</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb05</td>
<td>143</td>
<td>128.124</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb06, a</td>
<td>4</td>
<td>8.101</td>
<td>1.57</td>
</tr>
<tr>
<td>Sb06, b</td>
<td>84</td>
<td>104.490</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb07, a</td>
<td>19</td>
<td>42.729</td>
<td>0.261</td>
</tr>
<tr>
<td>Sb07, b</td>
<td>42</td>
<td>21.578</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb07, c</td>
<td>13</td>
<td>30.022</td>
<td>0.212</td>
</tr>
<tr>
<td>Sb08</td>
<td>131</td>
<td>109.908</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb09</td>
<td>114</td>
<td>111.440</td>
<td>0.00</td>
</tr>
<tr>
<td>Sb10</td>
<td>126</td>
<td>135.230</td>
<td>0.00</td>
</tr>
<tr>
<td>Total (whole-genome)</td>
<td>1163</td>
<td>1085.157</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Plant materials for SNP discovery

Raw sequence data from the resequencing of the three *S. bicolor* accessions described by Zheng et al. (2011) were made available prior to publication by the Chinese Academy of Sciences, Institute of Botany (Beijing, China). These included the sequences of the successful US sweet sorghum variety Keller, the Chinese sweet sorghum variety E-Tian and the Chinese grain sorghum variety Ji-2731. These three genotypes, which were resequenced by Zheng et al. (2011) to approximately 12× coverage with Illumina 100 bp paired-end sequencing, were complemented by whole-genome resequencing of the sweet sorghum line SS79 and the grain sorghum line M71, both from southern Africa. SS79 is an advanced sweet sorghum inbred line derived from a plant collected in a farmer’s field in Limpopo, South Africa, while the grain sorghum inbred line M71 (Macia SA) originates from an ICRISAT breeding programme in Zimbabwe (Shiringani et al., 2010). These two lines show considerable diversity with regard to plant height, photoperiodicity and flowering time, early-stage chilling tolerance, stem sugar and fibre, along with panicle and seed traits. Including the reference genome of the elite grain sorghum line Bt623 (Paterson, 2008; Paterson et al., 2009), our SNP discovery panel therefore consisted of three grain and three sweet sorghum genomes from diverse genetic and geographical origins.

Phenotype-based SNP discovery

In addition to the new whole-genome sequences of SS79 and M71, 60 F̂ RILs derived from a cross between SS79 and M71 were skim-sequenced at approximately 0.8× coverage. These 60 RILs were chosen by phenotypic screening to segregate for tolerance to prolonged early-season chilling stress, a key trait for European energy sorghum breeding.

Whole-genome sequencing of SS79, M71 and 60 SS79 x M71 RILs

High-quality, low-plastid, genomic DNA was extracted from *in vitro*-grown roots of SS79, M71 and the 60 selected RILs using DNeasy® Plant Mini Kits (QIAGEN GmbH, QIAGEN Strasse 1,
paired-end reads with insert sizes ranging from 180 to 440 nt were
40724 Hilden). DNA libraries were prepared using TruSeq
Illumina\textsuperscript{®} Paired-End DNA sample preparation kits (Illumina, Inc., San
Diego, CA), which generate DNA fragments from 200 to 500 nt in length, followed by 95 bp paired-end sequencing.
SS79 and M71 were each sequenced on one lane of an Illumina
Genome Analyzer Ix at Cologne Centre for Genomics (CCG),
Cologne, Germany. Raw sequences were quality-filtered, trimmed
and paired using the following parameters: Ambiguous = trim;
Maximum number of nucleotides in reads = 1000; minimum
number of nucleotides (nt) in reads = 15; Quality score = Illumina
Pipeline 1.5 and later.
The phenotype-based sequencing of 60 RILs was undertaken at
Beijing Genome Institute (BGI, Shenzhen, China) using random
paired-end sequencing of nebulized and bar-coded fragments.
The sequencing run was performed on a HiSeq 2000 sequencer.
A total of 360 251 856 quality-trimmed reads were obtained
with an average read length of 89.96 and an average distance
between pairs of 470.87 nt (Table 1).

Sequence analysis and SNP detection
Paired-end reads with insert sizes ranging from 180 to 440 nt were
perfectly. The average insert length after mapping was 345 bp. The
reads (average 93 bp) were mapped to the sorghum reference 79
accessed from Phytozome v9.0 (Goodstein et al., 2012), using
the CLC Genomics Workbench Version 4.9 (CLC Bio Aarhus, Denmark)
set to the following mapping parameters: Similarity = 0.8, length
fraction = 0.5, insertion cost = 3, deletion cost = 3, mismatch
cost = 2, global alignment = yes. Inclusion of broken reads is not
expected to change the accuracy of mapping (Li and Homer, 2010).
The conflict resolution for consensus sequence formation was
‘vote’, that is, the more frequent nucleotide will be reported,
however conflicts will be noted in downstream analyses such as
SNP detection. Raw sequences obtained from Chinese Academy
of Science were also filtered and mapped following the same pipeline.
Due to the higher depth of coverage (12 ×), in this case, we used
the ‘join maps’ function of the software CLC Genomics Workbench
version 4.9 after mapping the reads from each genotype into three
subgroups. CLC Genomics Workbench uses a neighbourhood
quality score (NQS) method for SNP calling (Altshuler et al., 2000).
We applied the following parameters: Window length = 11 nt,
maximum gap and mismatch = 2, central quality = 20 and aver-
age quality = 15. A SNP was called when it was detected in a
Figure 7  Colinearity between genetic and physical distance single nucleotide polymorphism (SNP) maps generated from the SM-RIL population (SS79 x M71). Blue circles indicate map positions calculated by a linear regression map order function, while the green circles show the map order results from maximum likelihood mapping. The latter resulted in a considerably improved fit of SNP marker orders to the expected positions within the sorghum genome.

SNP calling in phenotype-based RIL pools

Single-nucleotide polymorphism calling in the phenotype-based RIL pools was performed in two repetitions, once with highly stringent calling parameters and once with relaxed parameters. For the stringent calling, the minimum sequence coverage for the SNP detection in the two RIL pools was set to 10 reads, while a minimum SNP frequency of 99% was set for each group to ensure that only one allele per pool was permitted to be detected. Single-nucleotide polymorphism identification was subsequently performed once again using relaxed criteria, with a minimum coverage of four reads and a minimum SNP frequency of 2.4%, respectively. Single-nucleotide polymorphisms that were detected under stringent conditions at high frequencies in one group, and where the opposing allele was not detected under relaxed stringency in the second group, were considered to be unique to the pool in which they were found. Gene ontology (GO) term enrichment analysis of the genic SNPs was performed using the agrigo online analysis toolkit and database (Du et al. 2010).

SNP selection and array characterization

Candidate SNPs from the whole-genome data were filtered and selected by applying stringent criteria for genotype quality, allele frequency, the Illumina Assay Design Tool (ADT) score and the distribution throughout the genome. Besides the genome-wide SNPs, further SNPs were also selected in chromosome regions enriched with QTL for early-stage chilling tolerance, using the pooled sequence data. In this case, SNPs associated with the chilling-resistant genotype pool were preferentially selected based on proximity to potential candidate genes for abiotic stress tolerance and possibility of developing Infinium Type II assays. The Infinium Type II assay requires only one probe to detect both alleles for the most common SNP types (A/G, A/C, T/G, T/C), in contrast to the less common type I SNPs (A/T and C/G), which require two probes per SNP.

Figure 1 gives an overview of the scheme used for the SNP selection from the genome-wide SNPs and the SNPs in the phenotypic pools. The genome-wide SNPs were selected for even distribution throughout the entire genome, with an intentional absence of bias towards gene-rich regions or coding sequences. These SNPs were complemented by trait-linked SNPs showing polymorphisms between the phenotypic pools for early stage chilling tolerance. Finally, the SNP density was increased in important QTL-rich regions identified from multitrait phenotyping of the SS79 x M71 RIL population (unpublished data).

SNP array screening and genotype scoring

High-quality DNA samples from the 564 genotypes of the screening population and the control genotype (SS79) were extracted using a CTAB extraction protocol modified from Doyle et al. (2013) Society for Experimental Biology, Association of Applied Biologists and John Wiley & Sons Ltd, Plant Biotechnology Journal, 11, 1112-1125.
and Doyle (1990). A total of 92 RILs from the cross SS79xM71 were used to validate the SNPs. Furthermore, a diversity panel comprising 200 Sorghum spp. Accessions, maintained by KWS Saat AG (Einbeck, Germany), and three genetically divergent segregating populations (one F2, two F3) were also genotyped using the array (Table 3).

Microarray-based DNA genotyping was performed by Service-XS B.V., Leiden, the Netherlands, using the custom-designed iSelect BeadChip (Illumina, Inc., San Diego, CA) interrogating 2620 Infinium Type II SNPs. For each sample, 4 μL genomic DNA at ~50 ng/μL was processed and hybridized to the BeadChips, according to the manufacturer’s instructions. The BeadChip images were scanned on an Illumina iScan array reader, and the raw data were extracted into the Illumina GenomeStudio software (version 2011.1) using the default analysis settings. Genotyping analysis was performed using the Genotyping Module version 1.9.4 with the recommended default settings. A GenCall cut-off of 0.15 was applied, and clustering algorithm 2.0 was used. Clustering was performed using all samples and positive controls (with omission of the NTCs). Reproducibility error rates were calculated between the control sample replicates in each of the six screened plates.

The polymorphic information content (PIC) of the markers, which describes the measure of genetic diversity at a marker level dependent on the number of alleles and the frequency in a given population, was calculated in the 200 lines of the diversity panel plus the 8 mapping parents according to Anderson et al. (1993) using the following

\[
\text{PIC}_i = 1 - \sum_j p_{ij}^2
\]

where \(p_{ij}\) is the frequency of the \(j\)th allele for the \(i\)th SNP marker.

**Application of the array for population genetics analysis**

Population structure was analysed using the programme STRUCTURE 2.3.1 (Pritchard et al., 2000) using the 200 lines from the diversity collection plus the eight parents of the four mapping populations. The analysis was set at a burn-in period of 10 000 and 10 000 Markov chain Monte Carlo iterations, with 20 iterations to test \(K\) from 1 to 20 with 10 iterations for each \(K\) group. The optimum \(K\) value was calculated using the \(\Delta K\) system according to Evanno et al. (2005) using the online tool STRUCTURE harvester (Earl and vonHoldt, 2012).

**Linkage disequilibrium**

Linkage disequilibrium analysis was performed using the software TASSEL 4 (Bradbury et al., 2007) in two subsets of the diversity panel identified by STRUCTURE analysis as well as the entire diversity set (\(n = 208\)). The LD was calculated for all pairs of markers, and subsequent dissection of the LD estimates into chromosomes was then used to calculate LD (\(R^2\)) in 5 Mbp bins based on the physical positions of the markers. Decay in LD was calculated using the modified R code LDit.r (Ross-Ibarra Lab, University of Davis, CA, USA; see http://www.plantsciences.ucdavis.edu/faculty/ross-ibarra/code/files/LDit.html), which uses equation 1 from Remington et al. (2001) to estimate \(C\) employing nonlinear least squares and then plot the decay.

The critical LD value was calculated by square root transformation of the \(R^2\) values of the unlinked chromosome LD values and calculating the parametric 95th percentile according to Bresgeghello and Sorrells (2006). Linkage disequilibrium above this critical value is considered linked, and the interception point where the nonlinear regression model line meets this critical line is assumed to represent the population LD decay point.

**Linkage mapping**

Genetic linkage mapping was performed using the SNP calls from the 92 F2 RILs from the RIL population SS79 x M71 (SM-RIL), applying Haldane’s mapping function in the software JoinMap™ 4 (Kyazma, Wageningen, the Netherlands). The markers were grouped into linkage groups at LOD values from 8 to 15. Markers were ordered alternatively using regression and maximum likelihood methods, and the fit of marker orders for each mapping algorithm was compared with the sorghum physical map in relation to the projected positions of the SNP markers.

**Acknowledgements**

This work was partially funded by grant number 0315421B from the German Federal Ministry of Education and Research (BMBF), with additional support from the Federal Ministry for Consumer Affairs, Nutrition and Agriculture (BMVEL), Grant 23/12-13C HN12, and DAAD. The authors thank Hai-Chun Jing, Institute of Botany, Chinese Academy of Science, Beijing, China, for the provision of prepublication raw genome sequence data, Korbinian Schneeberger, Max-Planck Institute for Plant Breeding Research, Cologne, Germany, for assistance with SNP calling by SHOREmap, Wilbert van Workum (Service XS, Leiden, Netherlands) for generation of the SNP array and Janine Altmüller (CLC Genomics, Cologne, Germany) for generation of the genomic resequencing data.

**References**


Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 (a) Delta K (\(\Delta K\)) analysis plot to determine the most likely population substructure within a set of 200 diverse Sorghum bicolor accessions analysed for genetic diversity with the 3k SNP Infinium array described in this study. The clear peak for \(\Delta K\) at \(K = 2\) suggests the division of the population into two major subpopulations. (b) Graphical representation of population substructure within a set of 208 diverse Sorghum bicolor accessions analysed for genetic diversity with the 3k SNP Infinium array described in this study. Two clear subpopulations were detected.

Table S1 Summary of sorghum genotypes screened using the 3k SNP Infinium SNP array.

Data S1 Excel spreadsheet containing flanking sequence information, assay design scores, genome positions, minor allele frequencies and call rates for 3000 sorghum SNPs used to design the 3k SNP Infinium array described in this study.