Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor)

Zheng et al.
Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor)

Lei-Ying Zheng1, Xiao-Sen Guo2†, Bing He2†, Lian-Jun Sun1†, Yao Peng2, Shan-Shan Dong2, Teng-Fei Liu2, Shuye Jiang1,3, Srinivasan Ramachandran1,3, Chun-Ming Liu1 and Hai-Chun Jing1*

Abstract

Background: Sorghum (Sorghum bicolor) is globally produced as a source of food, feed, fiber and fuel. Grain and sweet sorghums differ in a number of important traits, including stem sugar and juice accumulation, plant height as well as grain and biomass production. The first whole genome sequence of a grain sorghum is available, but additional genome sequences are required to study genome-wide and intraspecific variation for dissecting the genetic basis of these important traits and for tailor-designed breeding of this important C4 crop.

Results: We resequenced two sweet and one grain sorghum inbred lines, and identified a set of nearly 1,500 genes differentiating sweet and grain sorghum. These genes fall into ten major metabolic pathways involved in sugar and starch metabolisms, lignin and coumarin biosynthesis, nucleic acid metabolism, stress responses and DNA damage repair. In addition, we uncovered 1,057,018 SNPs, 99,948 indels of 1 to 10 bp in length and 16,487 presence/absence variations as well as 17,111 copy number variations. The majority of the large-effect SNPs, indels and presence/absence variations resided in the genes containing leucine rich repeats, PPR repeats and disease resistance R genes possessing diverse biological functions or under diversifying selection, but were absent in genes that are essential for life.

Conclusions: This is a first report of the identification of genome-wide patterns of genetic variation in sorghum. High-density SNP and indel markers reported here will be a valuable resource for future gene-phenotype studies and the molecular breeding of this important crop and related species.

Background

Sorghum (Sorghum bicolor) originated from Africa and is a pro-poor multipurpose crop providing food, feed, fiber and fuel across a range of agro-ecosystems, especially in those with fragile conditions. Food and Agriculture Organization data show that sorghum is currently the number five most important grain crop and, in the past decade, its yearly production has been stabilized at 60 million tonnes with a harvesting area of 44 million hectares. Sorghum is known as ‘the camel amongst crops’ and requires much less water than many other cereals and has a remarkable ability to produce a crop under low levels of inputs and adverse stress conditions. Sweet sorghum is a natural variant of common grain sorghum with high stem sugar content and often considered a smart crop because it can produce both food and fuel. As a C4 crop with a high level of directly fermentable stem sugars and the ability to produce high biomass under adverse conditions, sweet sorghum is considered an ideal biofuel crop for the first and second generation bioethanol production, particularly having the advantages of exploitation of marginal land and avoiding competing for land for food crops [1-3]. However, the genetic basis for these remarkable traits of sweet sorghum is poorly understood.

Genetic variation consists of sequence variation and structure alteration. Sequence variation normally is manifested by SNPs, short sequence insertions and deletions (indels), microsatellites or simple sequence repeats, and transposable elements. The importance of SNPs and indels was initially realized by the occurrence of human sickle-cell anemia and cystic fibrosis diseases, the dramatic consequences caused by a nucleotide change in the hemoglobin beta gene [4] and a three-base deletion...
in the gene encoding a cystic fibrosis transmembrane conductance regulator [5-7], respectively. Structural alteration is generally described as presence/absence variations (PAVs) and copy number variations (CNVs), which include large scale deletions, insertions, duplications, inversions and translocations. An effect of CNV on phenotypic variation was documented 75 years ago in Drosophila melanogaster, with the Bar eye phenotype being caused by the Bar gene duplication [8,9]. In plants, sequence polymorphisms have gained much interest in the academic and breeding communities [10-12]. In several model and crop plants, including Arabidopsis, rice and maize, whole genome SNPs and indels have been developed [13-16] for a broad range of functional and evolutionary studies, including association mapping [17,18], genetic diversity [19,20], domestication, and genome evolution [21-23]. The effects of CNVs in plant genomes have only been reported in a few cases. In Arabidopsis and rice, array-comparative genome hybridization has been used to examine single-feature polymorphisms [24,25], genomic lesions caused by mutagenesis [26], as well as natural variation [27]. In maize, structural alterations have been reported to violate the intraspecific genome co-linearity [28,29] and contribute to the diversity of a range of important traits, such as heterosis and disease responses [22,30,31]. CNVs also shape the genome diversity of progeny of the immediate next generations in Arabidopsis [32]. Nonetheless, CNVs and their importance in plant genome and phenotypic variation are still far from well explored.

S. bicolor has three subspecies, namely arundinaceum, bicolor and drummondii, and the cultivated sorghums are all from bicolor, which has five local races, bicolor, caudatum, durra, guinea and kafir [33]. Although sweet sorghum differs phenotypically from grain sorghum and tends to have a sugar-rich juicy stem, taller plant, higher biomass but less grain production [1,34], how sweet sorghum differs genetically from grain sorghum well adapted to the northeast part of China with good seedling establishment and a short growth period?

These sorghum lines differ in a number of agronomic and biofuel-associated traits (Table 1). As expected, the two sweet sorghum lines (Keller and E-Tian) had taller

### Results

The morphological and physiological characteristics of sorghum lines used for resequencing

Sorghum (S. bicolor) accessions Keller, E-Tian, Ji2731 and the reference accession BTx623 were used for this work. Keller is an American-bred elite sweet sorghum line and has been shown to have good performance across a range of environmental conditions [41]. E-Tian (literally meaning Russian Sweet in Chinese) was a sweet sorghum line introduced to China in the early 1970s, while Ji2731 is a representative Chinese kaoliang grain sorghum well adapted to the northeast part of China with good seedling establishment and a short growth period.

These sorghum lines differ in a number of agronomic and biofuel-associated traits (Table 1). As expected, the two sweet sorghum lines (Keller and E-Tian) had taller
plant height, and higher stem Brix content and stem weight in comparison with the two grain sorghum lines. The Chinese sorghum line Ji2731 had zero accumulation of juice in the stem and the highest grain yield, which are the typical features of Chinese kaoliang. The variation in the biological traits in these four sorghum lines provides a basis to study gene-trait associations by examining the sequence polymorphisms and structural variations at the whole-genome level.

**Short-read resequencing and landscape of genome variation**

A whole-genome shotgun strategy and Illumina Genome Analyser sequencing technology were employed. The genome size of the reference genome BTx623 is 738,787,382, of which the effective size is 697,579,688 (excluding the N bases). We estimated that a $10 \times$ genome coverage should be sufficient for aligning most of the sequences. Nine paired-end sequencing libraries with an insert size around 500 bp, three for each sorghum line, were constructed using DNA samples from 10-day-old etiolated seedlings. Resequencing yielded 620.72 million 44-bp paired-end reads, which comprised 27.31 Gb of high-quality raw data. Sequence reads were aligned to the reference BTx623 genome using SOAP software v2.21 [42]. In total, we achieved an effective depth of $\times 36.51$ coverage, with an average of $\times 12.17$ for each line (Table S1 in Additional file 1).

With these reads and the information from the reference genome BTx623, including physical sequence alignment and gene models, we identified large quantities of SNPs, indels and PAVs (Figure 1). In total, 1,057,018 SNPs among these sorghum genomes, of which 83,262 SNPs were located in the coding regions, were identified (Table S2 in Additional file 1; Additional file 2). SOAPsnp [43] allows the detection of heterozygosity of SNPs and the results showed that the number of heterozygous SNP sites is less than 25% of all SNP sites over the whole genome or in coding regions of the sorghum genome (Table S3 in Additional file 1). The proportions of genic SNPs identified as coding, intronic, or UTR were 42.3%, 50.2%, and 7.5%, respectively. We also identified 99,948 indels ranging from 1 to 10 bp in length, of which 2,230 were in coding regions (Table S4 in Additional file 1).

The proportions of genic indels identified as coding, intronic, or UTR were 9.7%, 75.7%, and 14.6%, respectively. Moreover, 16,487 PAVs with an average length of 2,394 bp were identified (Table S5 in Additional file 1). Coding regions of 1,416 genes in sorghum genomes were included in these PAVs (Table S6 in Additional file 1). CNV was detected by using read depth of coverage [44]. A total of 17,111 CNVs, including 13,427 gains and 3,684 losses ranging from 2 kb to 48 Mb, were detected (5,994 for Ji2713, 3,603 for Keller and 7,514 for E-Tian).

Sanger sequencing technology was used for targeted gene verification. Primer sequences spanning genomic regions predicted to contain genetic variation were used to amplify genomic DNA templates from the three sorghum lines. In this manner, 215 SNPs in 30 genes were selected and 213 were verified using this method, suggesting a prediction accuracy of over 99% (Additional file 3). Similar accuracy was obtained with 48 indels and 9 CNVs.

Because some newly identified genes might exist beyond the currently assembled BTx623 sorghum genome, we assembled unmapped reads with SOAPdenovo and obtained contigs with a total sequence length of 7.2 Mb. Annotation of these contigs showed 73 putative absent genes with an average length of 409 bp (only coding regions were considered; Table S8 in Additional file 1). A Blast search against Arabidopsis, rice and maize genome databases revealed that 33 of these genes showed homology with known proteins (E-value $< 1e^{-6}$).

**SNP annotations and large-effect SNPs**

SNPs are small differences but with great impact on the variation of genomes and the biological traits. We therefore looked into the SNP annotations in detail and paid special attention to those in genic regions. For this purpose, the newly sequenced grain sorghum BTx623 genome was used as a reference [40]. Bearing in mind that all genome annotations, including that of sorghum, are imperfect and many factors affect the analysis of effects of SNPs - especially the presence of abundant transposon elements in the sorghum genome, which can be difficult to detect when they are present in low copy numbers, are even expressed, or contain fragments of ‘real’ genes - we analyzed the effects of SNPs using four different gene categories: bona fide genes, low-confidence genes, pseudogenes and transposons. We retrieved gene models of the Btx623 genome from the Phytozome database [45], and verified the gene identities using EST information from PlantGDB and maize gene orthologues from MaizeGDB [46]. In the end, the gene sets included 27,640 bona fide genes, 5,197 low-confidence genes, 932 transposons and 727 pseudogenes, respectively. These genes fall into 2,637 Pfam families (Additional file 4).

As shown in Table 2 the non-synonymous-to-synonymous ratios in the bona fide gene categories were the smallest, increasing from pseudogenes to transposons to low-confidence genes. It was also found that the bona fide Pfam-containing genes had a smaller ratio than those of the low-confidence genes and transposons (Table 3). Clearly, the presence of genes involved in transposon functions and transposases have strong effects on increasing the frequencies of SNPs in the genome and hence increasing the diversity of the genomes.
Figure 1 Genome-wide landscape of genetic variation in *Sorghum bicolor*. Gene density of chromosomes is visualized by line darkness; the more genes on a chromosome region, the darker the color. The purple and blue colors in the CNV ring represent gain and loss of copy number variation, respectively. For PAVs, the green color stands for the absence of variation, whereas pink for the presence of variation.

Table 2 Number and distribution of coding region SNPs in the resequenced sorghum genomes

| Sample  | Bona fide genes | Low confidence genes | Transposons | Pseudogenes | Total |
|---------|-----------------|----------------------|-------------|-------------|-------|
|         | Non-syn/Syn     | Non-syn/Non-syn/Syn | Non-syn/Non-syn/Syn | Non-syn/Non-syn/Syn |       |
| Ji2731  | 23,462          | 18,710               | 1.25        | 3,397       | 1.86  |
| Keller  | 14,091          | 10,346               | 1.36        | 2,048       | 2.02  |
| E-Tian  | 17,781          | 14,196               | 1.25        | 2,386       | 1.86  |
| Total   | 38,261          | 29,625               | 1.29        | 5,981       | 1.92  |

Syn, synonymous.
When the non-synonymous-to-synonymous ratios in the corresponding gene categories were compared, it was found that, overall, three out of the four gene categories displayed higher ratios in the coding regions of the genome, except for transposons (Tables 2 and 3), indicating that the Pfam domains possibly have fewer amino acid substitutions. A similar drop in the ratios was also reported for the rice genome [20], but the decrease in our sorghum genomes was smaller, which might be related to the imperfection of the genome annotations and the abundant presence of transposon elements.

We further analyzed the distribution of the SNPs in Pfam-containing genes in detail Figure 2. shows the number of non-synonymous and synonymous SNPs in individual Pfam gene families. The number of genes in the four gene categories is also indicated. Nearly half of the SNPs were found in leucine-rich repeats and genes encoding pentatricopeptide repeats (PPRs; Additional file 5), and another 20% of the total SNPs were from 10 Pfam family genes, including genes encoding protein kinases, protein tyrosine kinases, tetratricopeptide repeats, WD domain repeats, zinc knuckles, NB-ARC domains, ankyrin repeats, HEAT repeats, F-boxes, and armadillo/beta-catenin-like repeats. These Pfam family genes often have non-synonymous-to-synonymous ratios higher than 1, making the overall genome ratios high, and when these genes and the transposons were removed for the calculation, the ratio was reduced to close to 1 (data not shown). The finding that sequences encoding leucine-rich repeats and NB-ARC domains had higher ratios of non-synonymous to synonymous SNPs was consistent with findings in Arabidopsis, rice and maize, indicative of the diversification of plant disease-resistance proteins caused by pathogen pressure [19,20,22], whereas genes coding for ubiquitin, elongation factor Tu domain 2 and GTP binding domain proteins, all of which have important biological functions essential for life, had the lowest non-synonymous to synonymous ratios. Furthermore, we found that X8 domain and glycosyl hydrolase family 17 containing genes were amongst the families with the highest non-synonymous to synonymous ratios. As controls, the low-confidence genes, transposons and pseudogenes containing various Pfam domains were also included in the analysis. These genes tended to have more non-synonymous SNPs than synonymous SNPs, as represented by genes encoding BED zinc fingers, hAT family dimerization domains, and DUF domains, and MuDR family transposases. As these genes are not normally functional, caution should be exercised when interpreting the gene families with exceptionally high non-synonymous to synonymous ratios, which might not be bona fide genes, although evaluation of sorghum Pfam genes using the PlantGDB database showed that 71% of them had EST support.

We went further to analyze the distribution of so-called large-effect SNPs, which are predicted to have a potentially disabling effect on gene function. It was found that 1,664 SNPs were expected to induce premature stop codons, 65 to alter initiation methionine residues, 512 to disrupt splicing donor or acceptor sites, and an additional 16 to remove the annotated stop codons, resulting in longer open reading frames (Figure 3b). These large-effect SNPs are statistically significantly (P-value < 0.01) enriched in 14 Pfam families and depleted in 9 Pfam families (Figure 3a). However, large-effect SNPs were mostly enriched in transposase genes, such as those encoding MuDR family transposases, Transposase family tnp2 and retrotransposon gag proteins, or genes affected by transposon elements, such as hAT family dimerization domain and DUF domain genes. As these genes do not appear to be functional, we need to experimentally verify the importance of such enrichment. In contrast, those families devoid of large-effect SNPs, including ABC transporter and methyltransferase domain genes, are important for organism survival.

Taken together, we have identified large sets of SNPs, some of which are useful for further downstream functional genomics analyses if the SNPs reside in bona fide genes (SNPs residing in non-functional genes or transposon elements can also be useful as molecular markers). However, these results should be viewed with caution at this stage. As the identification of large-effect

### Table 3 Number and distribution of coding region SNPs in Pfam domain-containing genes in the resequenced sorghum genomes

|                | Bona fide genes | Low confidence genes | Transposons | Pseudogenes | Total |
|----------------|-----------------|----------------------|-------------|-------------|-------|
|                | Non-syn | Syn | Non-syn/Syn | Non-syn | Syn | Non-syn/Syn | Non-syn | Syn | Non-syn/Syn | Non-syn | Syn | Non-syn/Syn | Non-syn | Syn | Non-syn/Syn |
| J2731          | 17,694 | 14,798 | 1.20 | 921 | 567 | 1.62 | 489 | 259 | 1.89 | 0 | 0 | 0.00 | 897 | 504 | 1.78 |
| Keller         | 10,503 | 8,156 | 1.29 | 560 | 295 | 1.90 | 301 | 173 | 1.74 | 101 | 97 | 1.04 | 146 | 119 | 1.23 |
| E-Tian         | 13,439 | 11,197 | 1.20 | 600 | 373 | 1.61 | 341 | 219 | 1.56 | 146 | 119 | 1.23 | 71 | 60 | 1.18 |
| Total          | 28,551 | 23,182 | 1.23 | 1,608 | 965 | 1.67 | 897 | 504 | 1.78 | 209 | 174 | 1.20 | 1.26 |

Syn, synonymous.
sorghum genomes. The Pfam gene families with 30 or more non-synonymous and synonymous SNPs were analyzed and are listed. The Pfam genes are arranged according to the percentages of non-synonymous and synonymous SNP sites. The top Pfam gene families have lower percentages of non-synonymous SNP sites, while the bottom ones have higher percentages of non-synonymous SNP sites. The numbers in the non-synonymous and synonymous horizontal bars show the absolute numbers of SNPs, whereas the numbers in the gene categories are the numbers of genes in each category. For each Pfam, the number of genes in the categories of bona fide genes, low-confidence genes, transposons and pseudogenes are also listed. Gene numbers that are lower than 5% of the total genes analyzed are not shown. The chi-square significance of the observed non-synonymous and synonymous SNP distributions for each Pfam group is shown: *P-value < 0.05; **P-value < 0.001.

SNPs depends on the annotation of gene models, the exact number and spectrum of such SNPs will probably be modified when the genome annotation is updated. Furthermore, the sorghum lines used for genome resequencing differ in many traits, and the results may be biased towards having more large-effect SNPs. It would be interesting to include sorghum lines that closely phenotypically resemble each other for comparison. Finally, we analyzed only a limited number of sorghum lines; a representative collection of sorghum lines is required to justify the results obtained.

Effects of indels and presence/absence variations
We examined the genome-wide patterns of the 1- to 10-bp indels. With increasing indel size, the number of indels decreased. However, our result show that indels that are not multiples of 3 bp and produce frameshift mutations are particularly uncommon in coding regions but relatively common in non-coding regions (Figure 4a). We also found that genes with multiples of 3-bp indels were more commonly present in the genome than those with indels of other lengths (Figure 4b).

We further analyzed the Pfam domains affected by indels. Similar to the situation of non-synonymous SNPs, indels were statistically significantly (P-value < 0.001) enriched in NB-ARC and leucine rich repeat domain genes (Table S11 in Additional file 1). Indels were also found enriched in F-box, protein kinase and tyrosine kinase gene families, which are known to possess diverse functions and are suspected to follow a rapid birth-death cycle [47-49]. Although in theory the effects of frame-shifting (1-, 2-, 4-, 5-, 7-, 8- and 10-bp) indels are different from those of non-frame-shift (3-, 6- and 9-bp) ones, we found that, in sorghum, the gene families affected by them were very similar (Additional file 6). PAVs were highly enriched (P-value < 0.001) in nine gene families, some of which, such as NB-ARC.
**Figure 3** Annotation and distribution of SNPs

(a) Pfam gene families significantly (P-value < 0.01) enriched or depleted with large-effect SNPs. Asterisks indicate Pfam families statistically significantly depleted of large-effect SNPs.

(b) Statistics of different types of large-effect SNPs.

(c) Statistics of synonymous and non-synonymous SNPs.
domain genes, are the same as those enriched in SNPs and indels (Additional file 6).

**Genes with copy number variations**
The annotation showed that 2,600 genes had 3,234 CNVs, and 32 genes had CNVs in all three sorghum lines. Some of these 32 genes encode proteins involved in basic biological functions, such as RNA polymerase beta subunit (Sb04g009441, Sb04g009491, Sb02g017833), NADH dehydrogenase subunit 6 (Sb10g008595), ribosomal protein S7 (Sb05g020390) and ribosomal protein S18 (Sb02g032062). One gene, Sb04g035450, was found to be lost in Keller and E-Tian, but gained extra copies in Ji2731. A blast revealed that it is a homologue of the glutamate-gated kainate-type ion channel receptor subunit gene Glur5. Gene family enrichment analysis showed that CNVs were statistically significantly (P-value < 0.001) enriched in genes encoding cellulose synthases, pectinesterases, GRAS transcription factors, and BTB/POZ and auxin responsive proteins, in addition to the DUF, leucine-rich repeat and the zinc knuckle proteins (Additional file 2).

**Genetic variation between sweet and grain sorghum**
We speculated that some of the identified genetic variation might contribute to the phenotypic differentiation of sweet- and biofuel-associated traits and focused our analysis on SNPs, indels and PAVs in genic regions. For this, we used the gene set of the reference BTx623 genome as the control and identified all the shared variation within the two sweet sorghum lines and the variation between the reference genome and the Chinese local grain sorghum, respectively. Subsequently, the two sets of data were compared to remove those genes that have large-effect SNPs in the Chinese local grain sorghum, and the remaining gene set considered as those differentiating sweet and grain sorghum. We selected SNPs that are non-synonymous in the two sweet sorghum lines Keller and E-Tian but synonymous or even not present in the grain sorghum line Ji2731 (the reference BTx623 genome).
is a grain sorghum line). Similarly, for indels and PAVs, we selected those that were identified in both the Keller and E-Tian lines but not in Ji2731 and a gene was considered the putative differentiating gene when one indel was mapped to the coding region of the gene or the coding region of the gene was affected by a PAV.

Figure 5 shows the chromosomal locations of these genes, most of which are scattered in the vicinity of sub-telomeric regions, and obvious clusters of sweet-associated genes were found in genomic regions on chromosomes 4, 6 and 9. The selected SNP positions were mapped to 1,266 genes in this manner. A pathway enrichment analysis was performed for these genes, and as a result, ten Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, including starch and sucrose metabolism pathways and the lignin- and coumarine-biosynthesis associated phenylpropanoid biosynthesis pathways, were identified as statistically significantly (P-value < 0.1) enriched in the sweet-related gene sets (Table 4). In addition, 123 genes selected by indels and 53 genes selected by PAVs were identified. Among them, four genes were found to be in important gene families in previous studies. Sb10g024663 and Sb03g032210 were reported to be members of the P450 gene family, Sb09g017540 is a DREB transcription factor, and Sb01g034050 is in the expansin gene family. Owing to the limited grain and sweet sorghum genomes analyzed and the lack of a de novo-assembled complete sweet sorghum genome, the gene set reported here may not represent the whole spectrum of the genes differentiating sweet and grain sorghum and further in-depth study is required.

**Discussion**

Rapid development of sequencing technologies and bioinformatic tools makes the complete genome sequencing of many species possible, which provides a starting point to unravel the tremendous genetic variation and diversity at the genome scale. Amongst several model organisms examined to date, such as human, mouse, *Arabidopsis*, rice, and maize, genome-wide patterns of genetic variation are able to be captured by sampling a relatively small number of genomes [14,20,50-52]. By resequencing two sweet and one grain sorghum inbred lines, we uncovered nearly two million SNPs and indels, along with large numbers of PAVs and CNVs. This is a first report on the genome-wide patterns of genetic variation in sorghum, which will be valuable for further genotype-phenotype studies and for molecular breeding of this important *C₄* model crop.

Our study shows that the proportions of genic SNPs identified as in coding regions, intronic regions, or UTRs are 42.3%, 50.2%, and 7.5%, respectively. Compared to *Arabidopsis* [19] and rice [14,20], the intronic regions of sorghum genes harbor more SNPs. This might be related to the increased size of the introns; the average intron size for *Arabidopsis* is 168 bp, and for rice it is 397 bp, but for sorghum it is 444 bp. Our

![Figure 5 Chromosomal locations of genes differentiating sweet and grain sorghum](http://genomebiology.com/2011/12/11/R114)
results also demonstrate that, in sorghum, the proportions of large-effect SNPs resulting in premature stop codons, alteration of initiation methionine residues and disruption of splicing donor or acceptor sites are remarkably similar to what have been reported so far in Arabidopsis [19] and maize [22], but different from rice [20]. Furthermore, we found that 16 SNPs removed annotated stop codons and resulted in longer open reading frames, which is substantially smaller than the number (1,087) in maize.

It is known that transposon elements are abundant in sorghum as well as other cereal genomes [40,53]. As the genome annotation is not perfect, caution should be exercised with regard to the analysis of the effects of SNPs. Indeed, we found that the transposase genes, pseudogenes and low-confidence genes tended to have high non-synonymous-to-synonymous ratios in comparison with bona fide genes. This was reflected in the Pfam SNP annotations as well as in the analysis of so-called large-effect SNPs, which are predicted to disable gene functions. Most of the SNPs resided in receptor-like kinases, PPR repeats, disease resistant NB-ARC genes and other genes with multiple effects on stress responses. These genes also exhibited high non-synonymous-to-synonymous ratios, further supporting the notion from studies in other species that an arms race between plant-pathogen interactions results in diversification of the pathogen- or microbe-associated molecular pattern recognition receptors in plant genomes [54,55]. Significantly, the highest non-synonymous substitution ratios were found in X8 domain and glycoside hydrolase family 17 (glucan endo-1,3-beta-glucosidase) genes, which has not been reported in Arabidopsis [19], rice [20] or maize [22]. Current annotations show that limited low-confidence genes were included in these two Pfam gene families, although we cannot rule out the possibility that these genes are pseudogenes, or truncated because of the transposon elements. Further studies are required to validate whether they are related to specific biological processes in sorghum. However, the function of these genes in carbohydrate binding as well as in cell wall biosynthesis certainly provides clues to manipulating genes of interest for biofuel production.

In sorghum, the 14 gene families enriched with large-effect SNPs comprise genes encoding DUF proteins with unknown functions or include transposons, which appear to be nonfunctional but may affect genetic variation at the genome scale. Furthermore, gene families involved in biotic and abiotic stress tolerance, which do not contain transposons, also harbor enriched large-effect SNPs. For instance, over-expression of lecithin: cholesterol acyltransferase can increase lipid metabolism and the fluidity of membranes and hence the resistance to heat and/or cold shock (United States Patent Application 20050150007), whereas chalcone synthase in flavonoid biosynthesis and stilbene synthases for phytoalexin biosynthesis play important roles in sorghum disease resistance [56,57]. None of these gene families were reported to be enriched with large-effect SNPs in Arabidopsis, rice or maize. This could be due to genome/species-specific diversity, or result from the prediction algorithms used. Alternatively, this may also be related to the limited sorghum lines used, which have diverse relationships.

This effort also uncovered substantial numbers of indels and PAVs in the sorghum genomes. Indels that are not multiples of 3 bp were particularly uncommon in coding regions but relatively common in non-coding regions. This implies that most frameshift mutations are harmful to sorghum survival. The spectrum of gene families affected by indels and PAVs was similar to that of large-effect SNPs. This implies that although the origins and scales of affected genome segments may differ, SNPs, indels and PAVs may share similar survival and distribution patterns, at least in terms of gene families affected. CNV studies in plants lag behind those in animal and human models. Recent studies in maize showed its potential contribution to the heterosis of this crop during domestication and disease responses [22,30,31]. CNVs also shaped the genome diversity of progeny of the immediate next generations in Arabidopsis [32]. In the sorghum genomes, CNVs were present in several thousand genes, and some of the commonly involved genes are involved in basic biological functions as well as sugar- and bioenergy-associated traits. How this variation is associated with phenotypic variation is a new direction of future research.

The resequenced sorghum lines contained two elite sweet sorghum lines and one local elite Chinese grain sorghum line. We were able to identify genetic variation in 1,442 genes differentiating sweet and grain sorghum. Some of these genes are involved in the starch and sucrose metabolism pathway and the lignin- and

| KEGG ID   | P-value | Pathway name               |
|-----------|---------|----------------------------|
| sbi00230  | 0.058   | Purine metabolism          |
| sbi00500  | 0.077   | Starch and sucrose metabolism |
| sbi00240  | 0.017   | Pyrimidine                 |
| sbi00592  | 0.017   | alpha-Linolenic acid metabolism  |
| sbi00780  | 0.070   | Biotin metabolism          |
| sbi00940  | 0.030   | Phenylpropanoid biosynthesis |
| sbi04140  | 0.055   | Regulation of autophagy    |
| sbi04146  | 0.055   | Peroxisome                 |
| sbi03430  | 0.005   | Mismatch repair            |
| sbi03018  | 0.087   | RNA degradation            |
Sorghum is an important global crop, used for food, fodder, the production of alcoholic beverages, as well as biofuels. Genome-wide comparison studies with trait data of elite sorghum lines using the Joint Genome Institute Phytozome website [46]. Therefore genome sequences were downloaded from the Joint Genome Institute Phytozome website [46].

### Bioinformatics pipeline

**SNP detection**

We used a three-step procedure to detect high-quality SNPs. First, we calculated the likelihood of each accession's genotype using SOAPsnp [43]. Based on the alignment results, with consideration and analysis of data characteristics, sequencing quality and other experimental influences, the Bayesian model was applied to the actual data to calculate the probability of genotypes.

The genotype with the highest probability was selected as the genotype of the sequencing individual at the specific locus and a quality value was designated accordingly to reveal the accuracy of the genotype. Second, using the consensus sequence, a polymorphic locus against the reference sequence was selected. Third, on the basis of the resequencing data of three accessions, sites with sufficient quality, called effective sites, were used for SNP determination. Sufficient quality was based on the following criteria: $3 \leq \text{depth} \leq 50$, with depth calculated using data from each individual, and average mappable sites $< 1.5$. Candidate SNPs were those with sequencing depth of $3$ to $50$ for each accession and an average quality for the novel allele $> 20$. To exclude SNP calling errors caused by incorrect mapping or indels, we did not call two adjacent SNPs that were separated by $< 5$ bp. The remaining SNPs were defined as high quality SNPs. We performed SNP calling for each of the three accessions. These SNPs were used to calculate the whole genome SNP number and for

Conclusions

We report here a whole genome map of SNPs, indels, PAVs and CNVs amongst elite sorghum lines, which can be used as a framework for future comparative and functional genomics. Sorghum is an important global crop, used for food, fodder, the production of alcoholic beverages, as well as biofuels. Genome-wide comparison studies with trait data of elite sorghum lines using the SNPs, indels, PAVs and CNVs discovered here will provide additional clues to the molecular basis of the remarkable traits of sweet sorghum and will provide a powerful source for association genetics and discovery of alleles, which can be combined to achieve crop improvement in the future.

Materials and methods

**Plant materials and sequence data sets**

Four sorghum ($S.\ bicolor$) accessions were used in this study. BTx623 is an elite grain sorghum line used for whole-genome sequencing by the Joint Genome Institute and for making several mapping populations [38]. Keller (GRIN access code PI 653617) is an elite sweet sorghum line developed by DM Broadhead at US Sugar Crops Field Station at Meridan, Mississippi in 1982 and has been grown globally and proven to have good performance across a range of environmental conditions [41]. E-Tian (literally meaning Russian Sweet in Chinese) was a line introduced to China in the early 1970s and is known to have high stem Brix content, while Ji2731 is a Chinese local grain sorghum well adapted to the northeastern part of China with good seedling establishment and a short growth period (Professor Shi-Jie Gao and Dr Wei-Bin Gu, personal communication; Archives of Crop Varieties in Jilin Province, 1988).

Seeds were imbibed and germinated at 25°C in darkness under standard glasshouse conditions for 4 days and 10-day-old etiolated seedlings were harvested for DNA isolation using the CTAB (Hexadecyl trimethyl-ammonium bromide) buffer method. Following quality assessment, genomic DNA was randomly fragmented using sonification and size-fractionated through electrophoresis and DNA fragments of the desired length were gel purified. Adapter ligation and DNA cluster preparation were performed and subjected to Solexa sequencing according to the supplier’s protocol. The BTx623 reference genome sequences were downloaded from the Joint Genome Institute Phytozome website [46].

Bioinformatics pipeline

**SNP detection**

We used a three-step procedure to detect high-quality SNPs. First, we calculated the likelihood of each accession's genotype using SOAPsnp [43]. Based on the alignment results, with consideration and analysis of data characteristics, sequencing quality and other experimental influences, the Bayesian model was applied to the actual data to calculate the probability of genotypes. The genotype with the highest probability was selected as the genotype of the sequencing individual at the specific locus and a quality value was designated accordingly to reveal the accuracy of the genotype. Second, using the consensus sequence, a polymorphic locus against the reference sequence was selected. Third, on the basis of the resequencing data of three accessions, sites with sufficient quality, called effective sites, were used for SNP determination. Sufficient quality was based on the following criteria: $3 \leq \text{depth} \leq 50$, with depth calculated using data from each individual, and average mappable sites $< 1.5$. Candidate SNPs were those with sequencing depth of $3$ to $50$ for each accession and an average quality for the novel allele $> 20$. To exclude SNP calling errors caused by incorrect mapping or indels, we did not call two adjacent SNPs that were separated by $< 5$ bp. The remaining SNPs were defined as high quality SNPs. We performed SNP calling for each of the three accessions. These SNPs were used to calculate the whole genome SNP number and for
were kept as CNVs. The depth of each base from the alignment results; (ii) sequences were separated into fragments according to the orientation and appropriate span. If the orientation or span of the two paired-end reads is different from that of the gapped reads. For quality control, the final list of indels included only those identified on both strands by paired-end reads.

Presence/absence variation detection
According to the principal of paired-end sequencing, one of the paired-end reads should normally be aligned onto the forward sequence, while the other should be aligned onto the reverse sequence. The distance between the two aligned positions at the reference should be in accordance with the insert size. Thus, two paired-end reads aligned to the genome should have a normal orientation and appropriate span. If the orientation or span of the two paired-end reads is different from expectation for the alignments results, the region might then have structure variation. The abnormal paired-end alignments are analyzed by clustering and compared with the types of structure variation previously defined. In this manner, PAVs could be detected, with support from at least three abnormal paired-end reads. In this study, PAVs that were supported by at least six paired-end reads were thought to be of high quality and selected as the final PAVs.

Copy number variation detection
We detected CNVs by the following steps: (i) DNA sequences were separated into fragments according to the depth of each base from the alignment results; (ii) we calculated the P-value for each fragment to estimate its probability to be a CNV; and (iii) fragments that passed the criteria (fragment length longer than 2 kb, P-value ≤ 0.35, mean depth less than 0.5 or more than 2.0) were kept as CNVs. The P-value was calculated as the probability of each observed depth (d) under the distribution of a simulated Poisson distributed data set whose expected value (E(d)) equals the observed mean depth. If d < E(d), the P-value = P(x ≤ d) × 2, else P-value = P(x ≥ d) × 2. The credibility of a CNV increases as the P-value becomes smaller.
Additional file 2: Figure S1
Additional file 3: Table S7
Additional file 4: Table S9
Additional file 5: Table S10
Additional file 6: Figure S2

Abbreviations
Bp: base pairs; CNV: copy number variation; EST: expressed sequence tag; indel: insertion and deletion; KEGG: Kyoto Encyclopedia of Genes and Genomes; PAV: presence/absence variation; PPR: pentatricopeptide repeat; SNP: single nucleotide polymorphism; UTR: untranslated region.

Acknowledgements
We thank Professors Dongyun Hao and Shi-Jie Gao and Dr Wei-Bin Gu for the seeds and information of the sorghum lines used for resequencing. We also thank other members of Hai-Chun Jing’s lab and the bioinformatics team at BGI-Shenzhen for assistance in analyzing and experimentally verifying the data. Drs Kang Chong, Keping Ma, Sanwen Huang and Jun Wang are acknowledged for stimulating discussions during the initiation phase of this study. This work was funded by the grant of the Knowledge Innovation Programme (KIP) of the Chinese Academy of Sciences to Hai-Chun Jing (KSCQ-2-1M-G-067).

Authors details
1Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China.
2Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen 518083, China.
3Temasek Life Sciences Laboratory Limited, 1 Research Link National University of Singapore, Singapore 117604.

Authors’ contributions
SR, CML, and HCJ conceived and designed the experiments. HCJ analyzed the data; XSG, SR, CML, and HCJ contributed reagents/materials/analysis tools. XSG, BH, LJS, YP, SSD, TFL, SJ, SR, and HCJ performed the experiments. XSG, BH, LJS, YP, SSD, TFL, SJ, SR, and HCJ wrote the paper.

Competing interests
The authors declare that they have no competing interests.

Received: 11 April 2011 Revised: 4 November 2011 Accepted: 21 November 2011

References
1. Rooney WL, Blumenthal J, Bean B, Mullet JE: Designing sorghum as a dedicated bioenergy feedstock. *Biotechnol Biofuels* 2007, 1:4-17.
2. Carpira NC, McCann MC: Maize and sorghum: genetic resources for bioenergy grasses. *Trends Plant Sci* 2008, 13:415-420.
3. Vermeirssen W: Survey of genomics approaches to improve bioenergy traits in maize, sorghum and sugarcane. *J Integr Plant Biol* 2011, 53:105-119.
4. Ashley-Koch A, Yang Q, Olney RS: Sickle hemoglobin (Hbs) allele and sickle cell disease: a HuGE review. *Am J Epidemiol* 2000, 151:839-845.
5. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TA, Chakravarti A, Buchwald M, Tsui LC: Identification of the cystic fibrosis gene: genetic analysis. *Science* 1989, 245:1073-1080.
6. Ioannidis JP, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zelenksy J, Lok S, Plavsic N, Chou JL, et al: Identification of the cystic fibrosis gene: cloning and characterization of DNA. *Science* 1989, 245:1066-1073.
7. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Milmier G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaoka N, et al: Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989, 245:1059-1065.
8. Muller HJ: Bar Duplication. *Science* 1936, 83:528-530.
9. Bridges CB: The bar "gene" a duplication. *Science* 1936, 83:210-211.
10. Collard BC, Mackill DJ: Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 2008, 363:557-572.
11. Granov M, Altmann T, Roder MS: SNP identification in crop plants. *Curr Opin Plant Biol* 2009, 12:211-217.
12. Langridge P, Fleury D: Making the most of ‘omics’ for crop breeding. *Trends Biotechnol* 2011, 29:33-40.
13. Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weisshaar B: Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome Res* 2003, 13:1250-1257.
14. Foels FA, Wan J, Schulze SR, Eistl JC, Jiang N, Paterson AH: An SNP resource for rice genetics and breeding based on subspecies Indica and Japonica genome alignments. *Genome Res* 2004, 14:1812-1819.
15. Osowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D: Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res* 2008, 18:2024-2033.
16. Gore MA, Chia JM, Elshire RJ, Sun Q, Ehsis ER, Hurwit BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES: A first-generation haplotype map of maize. *Science* 2009, 326:1115-1117.
17. Huang X, Wen X, Tang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Deng L, Li W, Lu Y, Wang Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Li N, Buckler ES, Qian Q, Zhang QF, Li J, Han B: Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 2010, 42:961-967.
18. Kump KL, Bradbury PJ, Wissler RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zvonotcic JK, Kreovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB: Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 2011, 43:163-168.
19. Clark RM, Schweikert G, Toomajian C, Osowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schorkopf B, Nordborg M, Råtsh G, Ecker JR, Weigel D: Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 2007, 317:338-342.
20. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Bueler CR, Leung H, Leach JE: Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci USA* 2009, 106:12279-12284.
21. Wang L, Hao L, Li X, Hu S, Ge S, Yu J: SNP deserts of Asian cultivated rice: genomic regions under domestication. *J Evol Biol* 2009, 22:751-761.
22. Liu J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Liu D, Guo Y, Wang A, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J: Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 2010, 42:1027-1030.
23. van Heerwaarden J, Doebel J, Briggs WH, Glazebrock JC, Goodman MM, de Jesus Sanchez Gonzalez J, Ross-Ibarra J: Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc Natl Acad Sci USA* 2011, 108:1098-1092.
24. Borevitz JO, Liang D, Prouffe D, Zhang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J: Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* 2003, 13:513-523.
25. Kumar R, Qiu J, Yoshi T, Vallyodan B, Xu D, Nguyen HT: Single feature polymorphism discovery in rice. *PLoS One* 2007, 2:e284.
26. Bruce M, Hess A, Bai J, Mauleon R, Diaz MG, Sugiyama N, Bordeos A, Wang GL, Leung H, Leach JE: Detection of genomic deletions in rice using oligonucleotide microarrays. *J Genet Genomics* 2009, 36:1250-1257.
27. Salathia N, Lee HN, Sangster TA, Moronea K, Landry CR, Schellenberg K, Behere AS, Gunderson KL, Cavaleri D, Jander G, Queitsch C: Arabidopsis NABY: an indel array: an affordable alternative for genotyping. *Plant J* 2007, 51:727-737.
28. Fu H, Dooner HK: Intraspecific variation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* 2002, 99:9573-9578.
29. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A: Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 2005, 17:1483-1490.
30. Springer NM, Ying K, Fu Y, A T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Ingele AZ, Babarzak WB, Jeddeloh JA, Nettleton D, Schnable PS: Maize inbreds exhibit high levels of copy number variation.
52. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Jackson SM, Gillam B et al. The 873 maize genome: complexity, diversity, and dynamics. Science 2009, 326:1112-1115.

53. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pascenk S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reilly AD, Courney L, Krukowski SS, Tomlinson C, Strong C, DeLeauhyta K, Fronick C, Courney B, Rock SM, Beler E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jiang S, Jackson SM, Gillam B et al. The Sb73 sorghum genome: ancestry, adaptation, and dynamics. Science 2009, 326:457-463.

54. Zheng LY, Guo XS, He B, Sun LJ, Peng Y, Dong SS, Liu TF, Jiang S, Ramachandran S, Liu CM, Jing H-C. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience 2011 [doi:10.1186/1477-5958-10-12].

Cite this article as: Zheng et al.: Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biology 2011 12:R114.