Sorghum bicolor originated from Africa and became an important cereal crop after a long period of domestication and selective breeding [1]. Nowadays, it feeds over 500 million people in 98 countries [2], with an estimation of 42 million hectares of cultivated area and 62 million tons of yield per year (FAOSTAT data 2013, http://faostat3.fao.org). In contrast to C₃ crops such as rice and wheat, sorghum has the C₄ photosynthetic pathway, which leads to higher photosynthetic efficiency under circumstances of intense light, high temperature and low water supply [2–4]. As such, sorghum has remarkable drought and heat tolerance, and can produce high yield and biomass in areas of harsh conditions with low inputs. Sorghum is not only used for food, but also cultivated with other important economic impacts for forage, sugars and biomass. Furthermore, in recent years

**Background**

Sorghum (Sorghum bicolor) originated from Africa and became an important cereal crop after a long period of domestication and selective breeding [1]. Nowadays, it feeds over 500 million people in 98 countries [2], with an estimation of 42 million hectares of cultivated area and 62 million tons of yield per year (FAOSTAT data 2013, http://faostat3.fao.org). In contrast to C₃ crops such as rice and wheat, sorghum has the C₄ photosynthetic pathway, which leads to higher photosynthetic efficiency under circumstances of intense light, high temperature and low water supply [2–4]. As such, sorghum has remarkable drought and heat tolerance, and can produce high yield and biomass in areas of harsh conditions with low inputs. Sorghum is not only used for food, but also cultivated with other important economic impacts for forage, sugars and biomass. Furthermore, in recent years

**Abstract**

**Background:** Sorghum (Sorghum bicolor) is one of the most important cereal crops globally and a potential energy plant for biofuel production. In order to explore genetic gain for a range of important quantitative traits, such as drought and heat tolerance, grain yield, stem sugar accumulation, and biomass production, via the use of molecular breeding and genomic selection strategies, knowledge of the available genetic variation and the underlying sequence polymorphisms, is required.

**Results:** Based on the assembled and annotated genome sequences of Sorghum bicolor (v2.1) and the recently published sorghum re-sequencing data, ~62.9 M SNPs were identified among 48 sorghum accessions and included in a newly developed sorghum genome SNP database SorGSD (http://sorgsd.big.ac.cn). The diverse panel of 48 sorghum lines can be classified into four groups, improved varieties, landraces, wild and weedy sorghums, and a wild relative Sorghum propinuum. SorGSD has a web-based query interface to search or browse SNPs from individual accessions, or to compare SNPs among several lines. The query results can be visualized as text format in tables, or rendered as graphics in a genome browser. Users may find useful annotation from query results including type of SNPs such as synonymous or non-synonymous SNPs, start, stop of splice variants, chromosome locations, and links to the annotation on Phytozome (www.phytozome.net) sorghum genome database. In addition, general information related to sorghum research such as online sorghum resources and literature references can also be found on the website. All the SNP data and annotations can be freely download from the website.

**Conclusions:** SorGSD is a comprehensive web-portal providing a database of large-scale genome variation across all racial types of cultivated sorghum and wild relatives. It can serve as a bioinformatics platform for a range of genomics and molecular breeding activities for sorghum and for other C₄ grasses.

**Keywords:** Sorghum, Bio-energy plant, Genome variation, SNPs, Database curation
sorghum has been regarded as a promising bioenergy feedstock [5], which is comparable to other important biofuel grasses such as maize, sugarcane, Miscanthus and switch grass [6, 7]. Moreover, the compact genome and high degree of genetic synteny to other C₄ grasses make sorghum a potential genetic model for the design of bioenergy crops [8, 9].

Sorghum’s genome is relatively small (~730 M) and simple (10 chromosomes, diploid) compared to other C₄ crops in the Poaceae subfamily, such as maize and sugarcane. The recent completion and availability of a whole genome reference sequence, based on the elite line BTx623, has accelerated the pace of genetic and genomic research in sorghum [10]. The genetic basis of a range of important agronomic traits in sorghum has been elucidated, including drought tolerance and maturity [2]. Nevertheless, to better understand the genetic basis for the considerable phenotypic variation observed in many more agronomic and bioenergy traits of different sorghum accessions, it is necessary to have insight into genomic variation including single nucleotide polymorphisms (SNPs), insertions/deletions (INDELS) and structure variation (SV).

Recently, various high throughput strategies have been developed for genome re-sequencing [11–13], resulting in a large amount of SNP data being generated for sorghum [14–18]. These SNP data, representing high density biomarkers, are a valuable resource for researchers to perform genetic and breeding studies, such as genotyping by sequencing (GBS) [19–21], bulked segregant analysis (BSA) [22], and genome-wide association studies (GWAS) [18, 23, 24]. These studies will not only lead to the highly efficient discovery of key QTLs or genes relevant to important traits, but also contribute to the understanding of the evolutionary relationship of cultivated and wild Sorghum species and subspecies.

To enhance the utility of sorghum SNP data, we developed a web-based large-scale genome variation database (SorGSD, http://sorgsd.big.ac.cn). SorGSD contains ~62.9 million SNPs identified from the re-sequencing data of 48 sorghum lines mapped to the reference genome BTx623. These sorghum lines represent major cultivated races grouped into landraces or improved varieties, and weedy or wild subspecies. Figure 1 shows the phylogenetic relationship among these sorghum lines [16], with the genotype name and group indicated. Racial type and geographic origin are also included. Additionally, the total number of SNPs identified per sample is indicated. The two margaritiferum cultivars (PI525695 M Margaritiferum Mali 1964025 and PI586430 M Margaritiferum Sierra Leone 1938008) are separated into a distinct group since they are highly divergent from other S. bicolor races (Fig. 1). Two samples of the allopatric Asian species Sorghum propinquum are clustered within a distant group as the outgroup.

The SNP numbers of each sample give an overview of the genomic difference between the reference genome BTx623 and individual genomes. Detailed information about distribution of SNPs in different genomic regions, including genic, intergenic, and intronic regions is provided (Table 1). For genic regions, SNPs found in specific positions such as start and stop codons, splice donor and acceptor sites are listed (Table 2).

All the SNP data shown in the two tables can be easily accessed either as statistical information through the Help page of the database, or through the user interface. The original data of sequencing short reads, the assembled sequence and the SNP data of each accession can be downloaded.

User interface
SorGSD offers three main functions (search, compare and browse), for users to search, display and retrieve the SNPs and their annotations.

The search function provides a user-friendly web interface to query SNP information. Users can search SNPs by specifying chromosomal co-ordinates or the locus ID. Users can also query SNPs based on their genotypes, and predicted variant effects. In addition, users can compare the SNPs between two and more sorghum lines. The query results can be shown as a formatted table which contains the information of ID, chromosome position, genomic location and predicted coding effects, 5′ and 3′ flanking sequences, reference and derived alleles,
Fig. 1 A dendrogram showing the phylogenetic relationships among the diverse set of sorghum lines. Each sample is labelled as follows; the genotype name, sample type (coded, as detailed below), racial type, geographic origin, and total number of SNPs identified. Sample type codes: I - improved variety, L - landrace, W - weedy or wild, M - magaritiferum, P - Sorghum propinquum. The sorghum reference genome BTx623 is shown in bold, sweet sorghums are in italic. (Adapted from Mace et al. [16] and redrawn using the tool “Display Newick Trees” under MEGA 6.0, SS79 was added based on the output results of the SNPhylo program [34] using the SNP data.)
Table 1 Distribution of SNPs in different genomic regions in 48 sorghum accessions

| Genotype         | Type     | Racial type | Geographic origin | Genotype Type | Racial type | Geographic origin | Geographic origin | Total SNP numbers |
|------------------|----------|-------------|-------------------|---------------|-------------|-------------------|-------------------|-------------------|
| BTx623           | I        | Complex     | USA               | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Malisor84-7      | I        | Complex     | Mali              | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| ICSV745          | I        | Complex     | India             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| EarlyHegari      | L        | Caudatum    | Sudan             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Cherekit         | L        | Caudatum    | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC103-14E        | L        | Guinea-Caudatum | South Africa     | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Macia            | I        | Caudatum    | Mozambique        | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC108-14E        | L        | Caudatum    | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC237-14E        | L        | Caudatum    | Sudan             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| B923296          | I        | Complex     | Australia         | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| B963676          | I        | Complex     | Australia         | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| M35-1            | L        | Durra       | India             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| R931945-2-2      | I        | Complex     | Australia         | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC170-6-8        | L        | Caudatum    | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC326-6          | L        | Caudatum-Bicolor | Ethiopia-USA    | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC56-14E         | L        | Caudatum-Nigricans | Sudan            | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC62-14E         | L        | Complex     | Kenya             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| E-Tian           | I        | Kafir       | China             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Rio              | I        | Kafir       | USA               | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SS79             | I        | Kafir       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| IS8525           | L        | Kafir       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Keller           | I        | Complex     | USA               | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| RTx7000          | I        | Kafir-Caudatum | USA             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| IS3614-2         | L        | Guinea      | Nigeria           | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Kapern669        | I        | Complex     | USA-Sudan         | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI563516         | I        | Durra-Caudatum | Mali             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| QL12             | I        | Complex     | Australia         | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| IS9710           | L        | Caudatum    | Sudan             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| KS115            | I        | Durra-Caudatum | USA             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Ji2731           | L        | Caudatum    | China             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| A4               | I        | Complex     | China             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| LR9198           | I        | Complex     | China             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| BTx642           | L        | Durra       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC35C-14E        | L        | Durra       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| SC23-14E         | L        | Durra       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Yik.solute       | L        | Durra       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| IBC/E-38432      | L        | Durra       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI585749         | L        | Durra-Bicolor | Mali             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI30272         | W        | Drummondii  | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Zengada          | W        | Weedy       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Kilo             | W        | Weedy       | Ethiopia          | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| Greenleaf        | W        | Weedy       | USA               | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI226096         | W        | Weedy       | Kenya             | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI525695         | M        | Margaritiferum | Mali           | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI586430         | M        | Margaritiferum | Sierra Leone     | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| PI300119         | W        | Verticilliflorum | South Africa     | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
| AustRCF317961    | W        | Verticilliflorum | Australia     | All           | Intergenic   | 5’ UTR            | Intrinsic         | Non-syn           | Syn               | 3’ UTR            |
respectively. SNPs from the stringent set identified by both pipelines (see description in “Methods” and Fig. 2 for details) are highlighted with a green background in the result page. The output of the query results can be downloaded as flat text or formatted tables for further investigation.

SorGSD also provides several data browsing functionalities under the “Browse” pull-down menu. The “Total SNPs” tab lists the SNP numbers on 10 chromosomes of all 48 accessions. Users can select a group, e.g. Landraces, to display the SNP numbers of these accessions within this group. Mouse-clicking these SNP numbers will bring up the list of SNPs of a specific accession. Given that the different location in genes such as coding regions, as well as the non-synonymous information are often of great interest for further study, the “Genic SNP” tab lists several submenus including “Coding SNP”, “Synonymous SNP”, and “Non-synonymous SNP” so that information can be tailored to user requirements.

The “Browse on Chromosome” tab leads to an interactive graphic window to visualize SNPs in a genome browser. Users can customize the visualization interface by selecting different data types, including SNPs, genes, transcripts, allele frequencies, and the SNP density information. Users can obtain a pie chart showing the allele frequency, SNP density in 300 kb windows size, related gene and transcript information.

Help information
SorGSD provides a help resource for users to better access the SNP data, as well as proving links to additional sorghum research related resources. The help menu provides a “How to” page, which gives a number of examples for users to learn how to search and compare target SNPs. For example, a step-by-step user-guide shows how to obtain non-synonymous SNPs in chromosome 1 of sweet sorghum E-Tian, and how to compare SNPs between sweet sorghum E-Tian and two grain sorghum J12731 and Keller. An FAQs page provides answers to a range of frequently asked questions not only about the content and usage of SorGSD but more broadly about sorghum genomics. Detailed information including software tools, parameters and data sources is presented in the “Pipeline” page. The “Statistics” page shows the SNP numbers distributed in different genomic regions (Table 1) and specific genic sites (Table 2). The “Data source” page shows the general information of 48 sorghum lines, including their geographic origins, and links to the US Germplasm Resources Information Network (http://www.ars-grin.gov).

The “About” tab contains several pages related to sorghum research. The Sorghum Genome page provides a brief introduction to the reference genome BTx623, including genome size and gene number. The Resource page provides links to online databases, research institutions, sorghum producers and handbooks. The help page lists selected recently published papers in the fields of sorghum genomics, genetics, QTLs, etc., with links to full lists in PubMed.

Conclusions and future directions
High coverage resequencing data from two previous sorghum studies [15, 16] were used to identify SNPs among 48 sorghum genotypes by combining three SNP calling tools and updating the SNPs datasets using the sorghum reference annotation (Version 2.1). In addition, we annotated the effect of SNP variants on genes of each sorghum accession. SorGSD has already received over two thousands of visits from more than 30 countries around the world since it went online a few months ago. During the review process of this manuscript, we were happy to know that a new website Sorghum Genomics (https://www.purdue.edu/sorghumgenomics) developed at Purdue University became available as a functional gene discovery platform.

We will improve the SNP calling pipeline and the annotation procedure to obtain more accurate SNP data and upload them into the database. Furthermore, we will include additional types of genome variation data detected by newly developed pipelines, including INDELs and copy number variations (CNVs). At the same time, we will improve the web interface especially in the search function and give more examples in the user guide to help novice users to access the database easily. We will add more analytical functionalities so that users can perform more analyses such as Blast search, sequence alignment and phylogenetic analysis.

SorGSD can serve as a bioinformatics platform to inform wet-lab experiments including biomarker

---

Table 1 continued

| Genotype     | Type | Racial type | Geographic origin | Total SNP numbers |
|--------------|------|-------------|-------------------|-------------------|
|              |      |             |                   | All | Intergenic | 5’ UTR | Intron | Non-syn | Syn | 3’ UTR |
| Sorpr369-1   | P    | Propinquum  | –                 | 5200279 | 3971685 | 58105 | 713492 | 124517  | 141591 | 163430 |
| Sorpr369-2   | P    | Propinquum  | –                 | 4993948 | 3794524 | 53315 | 704812 | 118631  | 135432 | 160696 |

1 improved variety, L landrace, W wild/weedy, M margaritiferum, P Sorghum propinquum

---

Luo et al. Biotechnol Biofuels (2016) 9:6
| Genotype            | Type     | Racial type | Geographic origin   | Start codon | Stop codon | Splice sites |
|---------------------|----------|-------------|---------------------|-------------|------------|-------------|
| BTx623              | I        | Complex     | USA                 | 0           | 0          | 0           |
| Malisor84-7         | I        | Complex     | Mali                | 0           | 0          | 0           |
| IC57/745            | I        | Complex     | India               | 0           | 0          | 0           |
| Early-Hegari        | L        | Caudatum    | Sudan               | 0           | 0          | 0           |
| Cherekit            | L        | Caudatum    | Ethiopia            | 0           | 0          | 0           |
| SC103-14E           | L        | Guinea-Caudatum | South Africa   | 0           | 0          | 0           |
| Macia               | I        | Caudatum    | Mozambique          | 0           | 0          | 0           |
| SC108-14E           | L        | Caudatum    | Ethiopia            | 0           | 0          | 0           |
| SC237-14E           | L        | Caudatum    | Sudan               | 0           | 0          | 0           |
| B952296             | I        | Complex     | Australia           | 0           | 0          | 0           |
| B956376             | I        | Complex     | Australia           | 0           | 0          | 0           |
| M35-1               | L        | Durra       | India               | 0           | 0          | 0           |
| R951945-2-2         | I        | Complex     | Australia           | 0           | 0          | 0           |
| SC170-6-8           | L        | Caudatum    | Ethiopia            | 0           | 0          | 0           |
| SC526-6             | L        | Caudatum-Bicolor | Ethiopia-USA | 0           | 0          | 0           |
| SC56-6              | L        | Caudatum-Nigeriens | Sudan     | 0           | 0          | 0           |
| E-Tian              | I        | Kafir       | China               | 0           | 0          | 0           |
| I392                | I        | Kafir       | USA                 | 0           | 0          | 0           |
| K5115               | I        | Durra-Caudatum | USA          | 0           | 0          | 0           |
| J2731               | L        | Caudatum    | China               | 0           | 0          | 0           |
| A14                 | I        | Complex     | China               | 0           | 0          | 0           |
| LRR998              | I        | Complex     | China               | 0           | 0          | 0           |
| BTx642              | L        | Durra       | Ethiopia            | 0           | 0          | 0           |
| SC35-6              | L        | Durra       | Ethiopia            | 0           | 0          | 0           |
| SC23-6              | L        | Durra       | Ethiopia            | 0           | 0          | 0           |
| Yir-solate          | L        | Durra       | Ethiopia            | 0           | 0          | 0           |
| IBC/E-384322        | L        | Durra       | Ethiopia            | 0           | 0          | 0           |
| PI585749            | L        | Durra-Bicolor | Mali          | 0           | 0          | 0           |
| PI30272             | W        | Drummondii  | Ethiopia            | 0           | 0          | 0           |
| Zengada             | W        | Weedy       | Ethiopia            | 0           | 0          | 0           |
| Kilo                | W        | Weedy       | USA                 | 0           | 0          | 0           |
| Greenleaf           | W        | Weedy       | USA                 | 0           | 0          | 0           |
| PI226096            | W        | Weedy       | Kenya               | 0           | 0          | 0           |
| PI526593            | M        | Margaritiferum | Mali          | 0           | 0          | 0           |
| PI586430            | M        | Margaritiferum | Sierra Leone     | 0           | 0          | 0           |
| PI300119            | W        | Verticilliflorum | South Africa | 0           | 0          | 0           |
| AustTRCF317961      | W        | Verticilliflorum | Australia | 0           | 0          | 0           |
development, allele mining and gene function assessment. In addition to the collaboration among research groups involving in this work, we will collaborate with other domestic and international laboratories in the sorghum research community to sequence and annotate more sorghum accessions in the future.

We will update the database regularly and add SNP datasets with newly available re-sequenced sorghum accessions. We hope that the high density of these SNP data at genomic level collected from the major races of cultivated sorghum as well as other subspecies is a rich repository for a broader research community working in biomarker identification, genetic analysis and molecular breeding, especially for energy plant sweet sorghum cultivation.

Methods

The construction of SorGSD was a multi-step process. Firstly, the sorghum re-sequencing paired-end raw reads reported in the previously published works were downloaded [15, 16]. In addition, the paired-end raw reads generated in-house for a sweet sorghum line SS79 were included [unpublished data]. Secondly, the raw reads were mapped to the reference sorghum genome (BTx623) [10] using the BWA program [25]. SNPs were identified using the software GATK [26, 27], realSFS (http://popgen.dk/angs/index.php/RealSFS) and SOAPsnp [28] and annotated using SnpEff [29]. With the SNP matrix finalized, a web interface was designed for users to browse and search the SNPs and related annotations. Details for the database construction are described as follows and are also available on the designated website.

Data source

The raw reads of sequencing data were from three original datasets. The largest dataset [16] contains 44 sorghum accessions and represent the major races of cultivated sorghum as well as their wild relatives. The second dataset [15] contains three accessions of cultivated sorghums. The raw reads of these two datasets can be downloaded from the NCBI sequence read archive (SRA) (accessions SRS378430-SRS378473, and accessions SRX100115-SRX100138). The third dataset contains the paired-end reads of sorghum line SS79, a cultivated sweet sorghum inbred. These data were recently generated in our laboratory using an Illumina HiSeq 2000 platform with insert size of 500 bp and have not been submitted to NCBI. The average sequencing depth of all sorghum accessions is about 20×, ranging from 12 to 54×.

SNP calling pipeline

After trimming adapters, the clean reads were mapped to version 2.1 of the reference genome (available via http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sobicolor) using the BWA program [25], allowing a maximum of five mismatches and disabling long gaps in the mapping procedure. The average counts of the mapping rate, the unique mapping rate and the mapping coverage were 0.957, 0.681 and 0.881 respectively, excluding the two S. propinquum accessions. The SAM tools package [30] was used to convert mapping results to BAM format, and then the Picard program (http://picard.sourceforge.net) was applied to eliminate duplicated reads generated during the process of library construction.
Subsequently, the GATK tools [26, 27] were used to recalibrate the base quality score to obtain more accurate quality scores for each base and realign reads around known INDELs. The refined data from all individuals were jointly used to call a raw SNPs set by GATK Haplotype-Caller. Finally, a set of SNPs were identified, using the variant quality score to recalibrate the procedure in GATK. In total, we identified 62,888,582 SNPs across all 48 sorghum lines, corresponding to 15,357,261 sites in the reference genome. The GATK based SNP calling pipeline is similar to that reported in a recent publication [31]. SNPs were additionally identified using the pipeline described previously using realSFS (http://popen.dk/angs/index.php/RealSFS) and SOAPsnp [28], described by Mace et al. [16]. Approximately 28 million highly stringent SNPs were in common between the two SNP identification pipelines (Fig. 2) with the GATK-based pipeline identifying more SNPs than the SOAPsnp-based pipeline. The total number of SNPs called by the GATK based pipeline was found to be comparable to the study by Evans et al. [32], which employed the CLC Workbench software (CLC Bio-Qiagen, Aarhus, Denmark). All the SNPs identified by the GATK pipeline were stored in SorGSD, with the subset of 28 million highly stringent SNPs highlighted in the results page. Finally, the effect of variants on all the v2.1 predicted gene models for each sorghum accession were predicted and annotated using the SnpEff program (version 4.0e) [29].

Database implementation

The SNP data and their related annotations were formatted into tables and stored in SorGSD using the MySQL database management system (version 5). The web interface of SorGSD was designed by JAVA/JSF (JDK 1.6) under the Apache/Tomcat web server (version 2.0) running under a Linux operation system (CentOS 6). We installed the generic genome browser GBrowse [33] as a chromosome-based visualization tool to display these genomic SNPs and annotations.

Abbreviations

SNP: single nucleotide polymorphism; INDEL: insertion/deletion; SV: structure variation; GBS: genotyping by sequencing; BSA: bulked segregant analysis; GWAS: genome-wide association study; QTLs: quantitative trait locus; CNV: copy number variation.

Authors’ contributions

HL and WMZ initiated the SorGSD project and designed the database structure. YQW and BXT constructed the database. WMZ, YQW and JZW maintain the web server. JYZ, HL and JCL designed the web interface. YX, HL, XYW, LMZ, LF, ZLD, WAB and SST participated in data analysis. DRJ, IDG, RJS, ESM and HCJ coordinated the sorghum SNP projects. HL drafted the manuscript. JCL, HCJ, ESM, IDG and RJS revised the manuscript. All authors read and approved the final manuscript.

Author details

1. Genomics and Molecular Breeding of Biofuel Crops, Key Laboratory of Plant Resources, Institute of Botany, Chinese Academy of Sciences, 100093 Beijing, China. 2. Beijing Institute of Genomics, Chinese Academy of Sciences, 100101 Beijing, China. 3. Department of Plant Breeding, Justus Liebig University, Giessen, Germany. 4. BGI-Shenzhen, 518083 Shenzhen, China. 5. Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Warwick, QLD 4370, Australia. 6. School of Agriculture and Food Sciences, The University of Queensland, Brisbane, QLD 4072, Australia. 7. Department of Agriculture, Fisheries & Forestry (DAFF), Warwick, QLD 4370, Australia. 8. College of Life Sciences and State Key Laboratory of Protein and Plant Gene Informatics, Wageningen University and Research Centre, Wageningen, The Netherlands.

Acknowledgements

Thanks to the members of our laboratories for their useful suggestions to improve the user interface of the database. We are grateful to the anonymous reviewers for their critical comments and suggestions to improve the web pages of the database.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by grants to Hai-Chun Jing from the National Natural Science Foundation of China (31461143023, 31271797), National Science and Technology Support Program (2015BAD15B03, 2013BAD22B01) and Sino-Africa Centre of CAS International Outreach Initiatives.

Received: 29 September 2015 Accepted: 10 December 2015

Published online: 07 January 2016

References

1. Doggett H. Yield increase from sorghum hybrids. Nature. 1967;216:798–9.
2. Pennisi E. Plant genetics: how sorghum withstands heat and drought. Science. 2009;323:573.
3. Osborne CP, Beerling DJ. Nature's green revolution: the remarkable evolutionary rise of C4 plants. Philos Trans R Soc Lond B Biol Sci. 2006;361:173–94.
4. Sasaki T, Antonio BA. Plant genomics: sorghum in sequence. Nature. 2009;457:547–8.
5. Rooney WL, Blumenthal J, Bean B, Mullet JE. Designing sorghum as a dedicated bioenergy feedstock. Biofuels, Bioprod Biorefin. 2007;1:147–57.
6. Carpenter NC, McCann MC. maize and sorghum: genetic resources for bioenergy grasses. Trends Plant Sci. 2008;13:415–20.
7. Vermoesen W. Survey of genomics approaches to improve bioenergy traits in maize, sorghum and sugarcane free access. J Integr Plant Biol. 2011;53:105–19.
8. Calvino M, Messing J. Sweet sorghum as a model system for bioenergy crops. Curr Opin Biotechnol. 2012;23:323–9.
9. Mullet J, Morishige D, McCormick R, Truong S, Hilley J, McKinley B, Anderson R, Olson SN, Rooney W. Energy sorghum—a genetic model for the design of C4 grass bioenergy crops. J Exp Bot. 2014;65:3479–89.
10. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457:551–6.
11. Eshri RR, Glabuzic JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SA. A Robust, Simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379.
12. Davey JW, Hohenlohe PA, Etter PO, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.
13. Wang S, Meyer E, McKay JK, Matz MV. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Meth. 2012;9:808–10.
14. Nelson JC, Wang S, Wu Y, Li X, Antony G, White FF, Yu J. Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genom. 2011;12:352.
15. Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M, Jing H-C. Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol. 2011;12:R114.
16. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, et al. Whole-genome sequencing reveals untapped genetic potential in Africa’s indigenous cereal crop sorghum. Nature Commun. 2013;4:2320.

17. Bekele WA, Wiekhorst S, Friedt W, Snowdon RJ. High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. Plant Biotechnol J. 2013;11:1112–25.

18. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci. 2012;110:4553–8.

19. Nielson R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.

20. Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor Appl Genet. 2013;126:2699–716.

21. Morishige D, Klein P, Hilley J, Sahraei SM, Sharma A, Mullet J. Digital genotyping of sorghum—a diverse plant species with a large repeat-rich genome. BMC Genom. 2013;14:448.

22. Han Y, Lv P, Hou S, Li S, Ji G, Ma X, Du R, Liu G. Combining next generation sequencing with bulked segregant analysis to fine map a stem moisture locus in sorghum [Sorghum bicolor (L.) Moench]. PLoS ONE. 2015;10:e0127065.

23. Rhodes DH, Hoffmann L, Rooney WL, Ramu P, Morris GP, Kresovich S. Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm. J Agric Food Chem. 2014;62:10916–27.

24. Adeyaju A, Little C, Yu J, Teso T. Genome-wide association study on resistance to stalk rot diseases in grain sorghum. G3 (Bethesda). 2015;5(6):1165–75.

25. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.

26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garmire L, Altshuler D, Daly M, DePristo MA. The genome analysis toolkit: a Mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

27. DePristo MA, Banks E, Poplin R, Gatlin KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

28. Li R, Li Y, Fang X, Yang H, Wang J, Kristansen K. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19:1124–32.

29. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.

30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence alignment/map format and SAM tools. Bioinformatics. 2009;25:2078–9.

31. McCormick RF, Truong SK, Mullet JE. RIG: recalibration and interrelation of genomic sequence data with the GATK. G3 (Bethesda). 2015;5:665–65.

32. Evans J, McCormick RF, Morshige D, Olson SN, Weers B, Hilley J, Klein P, Rooney W, Mullet J. Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. PLoS One. 2013;8:e79192.

33. Stein LD, Mungall C, Shu S, Caudy A, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arna A, Lewis S. The generic genome browser: a building block for a model organism system database. Genome Res. 2002;12:1599–600.

34. Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genom. 2014;15:162.