

Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential

Gengyun Zhang^{1,2,6}, Xin Liu^{1,6}, Zhiwu Quan^{1,2,6}, Shifeng Cheng^{1,6}, Xun Xu^{1-3,6}, Shengkai Pan^{1,6}, Min Xie¹, Peng Zeng¹, Zhen Yue¹, Wenliang Wang¹, Ye Tao¹, Chao Bian¹, Changlei Han¹, Qiuju Xia^{1,2}, Xiaohua Peng^{1,2}, Rui Cao¹, Xinhua Yang¹, Dongliang Zhan¹, Jingchu Hu¹, Yinxin Zhang^{1,2}, Henan Li^{1,2}, Hua Li^{1,2}, Ning Li^{1,2}, Junyi Wang¹, Chanchan Wang^{1,2}, Renyi Wang^{1,2}, Tao Guo^{1,2}, Yanjie Cai^{1,2}, Chengzhang Liu^{1,2}, Haitao Xiang^{1,2}, Qiuxiang Shi^{1,2}, Ping Huang^{1,2}, Qingchun Chen^{1,2}, Yingrui Li¹, Jun Wang^{1,4}, Zhihai Zhao⁵ & Jian Wang^{1,2}

Foxtail millet (*Setaria italica*), a member of the Poaceae grass family, is an important food and fodder crop in arid regions and has potential for use as a C₄ biofuel. It is a model system for other biofuel grasses, including switchgrass and pearl millet. We produced a draft genome (~423 Mb) anchored onto nine chromosomes and annotated 38,801 genes. Key chromosome reshuffling events were detected through collinearity identification between foxtail millet, rice and sorghum including two reshuffling events fusing rice chromosomes 7 and 9, 3 and 10 to foxtail millet chromosomes 2 and 9, respectively, that occurred after the divergence of foxtail millet and rice, and a single reshuffling event fusing rice chromosome 5 and 12 to foxtail millet chromosome 3 that occurred after the divergence of millet and sorghum. Rearrangements in the C₄ photosynthesis pathway were also identified.

Foxtail millet (*Setaria italica*) was domesticated >8,700 years ago¹. Although foxtail millet was an important foodstuff in neolithic China and continues to be cultivated in semi-arid areas, the specific growth conditions and low productivity have limited its potential as a crop. Foxtail millet is closely related to several biofuel grasses with complex genomes, such as switchgrass (*Panicum virgatum*), napier grass (*Pennisetum purpureum*) and pearl millet (*Pennisetum glaucum*). Of these related species, foxtail millet is the most suitable plant for whole genome sequencing and analysis owing to its small genome (~490 Mb)², self-pollination, rich genetic diversity (~6,000 varieties)^{3,4}, the existence of a complete collection of germplasm and the availability of high-throughput transformation platforms².

Genetic maps have already been constructed for foxtail millet using different markers^{3,4}. Comparisons between genetic maps of foxtail millet, rice and pearl millet have shed light on the chromosomal relationships between these grasses^{4,5}. Genetic variations of foxtail millet have also been studied in some genomic regions⁶ and population of foxtail millet has been investigated using single-nucleotide polymorphism (SNP), simple sequence repeat and other markers^{7,8}. Evolutionary relationships between foxtail millet and its wild ancestor green millet have been studied using polymorphisms in single genes⁹ or genomic regions⁶. Small collections of other resources, including simple sequence repeats¹⁰ and expressed sequence tags (ESTs)¹¹, have also been developed. The process of C₄ photosynthesis and genes in C₄ pathway has been investigated¹². The availability of genome

sequences has enabled the identification of gene duplication and neo-functionalization events that have contributed to the evolution of several enzymes involved in C₄ photosynthesis¹³. Carboxylase, which catalyses the first reaction in C₄ photosynthesis, is thought to differ in C₃ and C₄ plants¹⁴. However, evolution of carboxylase genes in C₃ and C₄ plants is still not well understood, especially in grasses. Using the foxtail millet genome draft, we were able to analyze carboxylase gene evolution in C₃ and C₄ grasses.

We used whole genome shotgun combined with next-generation sequencing to assemble a draft genome of the foxtail millet strain 'Zhang gu'. The final genome assembly was 423 Mb, ~86% of the estimated genome size, with repeats comprising ~46% of the genome. We annotated 38,801 protein-coding genes, of which ~81% were expressed. Nine foxtail millet chromosomes were found to be formed after three chromosomal reshuffling events. The carbonic anhydrase genes in the C₄ pathway were found to differ between C₃ and C₄ grasses. Finally, using developed structural variant and SNP markers, we mapped the setoxydim-resistance trait to a known gene¹⁵.

RESULTS

Sequencing and assembly

Whole genome shotgun combined with next-generation sequencing has been used to assemble large mammal¹⁶⁻¹⁸ and plant genomes¹⁹⁻²¹. Although this method has proven limitations when used to assemble long-repeat regions²², improvements in sequencing (using paired-end

¹BGI-Shenzhen, Shenzhen, China. ²BGI-Shenzhen, Chinese Ministry of Agriculture, Key Lab of Genomics, Shenzhen, China. ³BGI-Americas, Cambridge, Massachusetts, USA. ⁴Department of Biology, University of Copenhagen, Copenhagen, Denmark. ⁵Institute of Millet, Zhangjiakou Academy of Agricultural Science, Zhangjiakou, China. ⁶These authors contributed equally to this work. Correspondence should be addressed to Jun W. (wangj@genomics.org.cn), Z.Z. (zhaozhihai58@163.com) or Jian W. (wangjian@genomics.org.cn).

Received 12 December 2011; accepted 26 March 2012; published online 13 May 2012; doi:10.1038/nbt.2195

Table 1 Summary of genome assembly and annotation

		Assembly		
		N50 (size/number)	N90 (size/number)	Total length
Genome assembly	Contigs	25.4 kb/4,667	5.3 kb/16,903	394 Mb
	Scaffolds	1.0 Mb/136	258 kb/439	423 Mb
	Chromosomes	9 chromosomes (from 613 scaffolds)		400 Mb
		Annotation		
		Total length		
Transposable elements	Total	196.6 Mb (46.3%)		
	Retroelements	133.6 Mb (31.6%)		
	DNA transposons	39.7 Mb (9.4%)		
		Copies		
		Total length		
Noncoding RNAs	rRNAs	99	18.7 kb	
	tRNAs	704	52.8 kb	
	miRNAs	159	19.3 kb	
	snRNAs	382	43.6 kb	
Protein coding genes	Total number	Supported by transcriptome data	Homologous genes with sorghum	Function assigned
		38,01	31,709	32,701

sequencing) and bioinformatics (more efficient assemblers)^{23,24} have enabled resolution of repeat-rich regions from such data^{25,26}. We used the whole genome shotgun–next-generation sequencing strategy to assemble the draft genome of a widely cultivated strain of foxtail millet named Zhang gu. DNA libraries with different insert sizes ranging from 170 bp to 40 kb (**Supplementary Table 1**) were cloned to resolve sequences of repetitive regions. Sequencing was carried out using Illumina second-generation sequencers. After filtering, ~40 Gb data were subjected to SOAPdenovo²³ which used *de Bruijn* graphs for genome assembly using pair-end short-reads. After gap filling, the contig N50 was 25.4 kb, and 90% (in length) of the contigs were contained in 16,903 contigs (**Table 1** and **Supplementary Table 2**). The scaffold N50 was 1.0 Mb and 90% of the scaffolds (380 Mb) were contained in the 439 longest scaffolds. The total length of all scaffolds was 423 Mb, with 28 Mb (6.6%) gaps. The genome size was estimated to be ~490 Mb by cytogenetic methods², which was consistent with the estimation by k-mer analysis carried out in this study (**Supplementary Notes**) (~485 Mb). Thus the scaffolds covered ~86% (81% if excluding gaps) of the total genome.

Marker development and genetic map construction

We resequenced another widely cultivated strain of foxtail millet named 'A2' to identify variations and construct a genetic map for anchoring scaffolds onto chromosomes. A2 is a photothermo-sensitive male-sterile line. We used Illumina GA II to sequence the A2 strain to ~10× in depth. We identified 542,322 SNPs, 33,587 small insertion and deletions (indels), and 10,839 structural variants between A2 and Zhang gu (**Supplementary Notes**). Then we constructed an F₂ population with 480 individuals by crossing Zhang gu and A2. We genotyped 759 variations including 118 SNPs and 641 structural variants (**Supplementary Table 3**) in the F₂ population. Of those markers, the 751 (covering the 613 longest scaffolds) that showed segregation in the F₂ population were used to construct the genetic map. The 751 markers were clustered into nine linkage groups (**Supplementary Fig. 1**), and we were able to anchor the 613 longest scaffolds (~400 Mb, with ~26 Mb gaps) onto nine chromosomes.

We compared the genetic distance to the physical distance of these 751 genetic markers (**Supplementary Fig. 2**) and found that 33% of the genome was located in low-recombination regions, a percentage lower than in sorghum (62%)²⁷ but higher than in rice (15%)²⁸. The average ratio of genetic-to-physical distance in low- and high-recombination

regions was 0.44 cM/Mb and 6.77 cM/Mb, respectively. In the low-recombination regions, we identified a clustered 155-bp repeat unit, which was similar in sequence to the 140-bp centromere elements in sorghum²⁹. This repeat unit is likely to be a constitutive, centromere-related element.

Assessment of genome assembly

We first mapped all the individual reads back to the genome assembly (**Supplementary Table 4**) and found that >89% of the reads could be mapped, which was consistent with the estimated coverage of the assembly. Among mapped reads, ~8% were mapped without a proper paired-end relationship, which could be caused by difficulty in mapping or a discrepancy in assembly. We compared our genome assembly with the sorghum genome (version 1.0) to identify 1,937

large-scale rearrangements (**Supplementary Table 5**), and most of them (99%) were well supported by paired-end mapped reads; ~1% was not well supported, which might be because of assembly errors (**Supplementary Fig. 3**). We also used optical mapping³⁰ to validate our genome assembly. We mapped 99.8% of the optical fragments to the assembly (**Supplementary Fig. 4**), which confirms the accuracy of the genome assembly. Finally we used a different assembler, ALLPATHS-LG²⁴, to carry out a second assembly. The two assemblies were comparable (**Supplementary Tables 2** and **6**), although the assembly by ALLPATHS-LG had longer contigs, longer scaffolds, more gaps and shorter total length. We mapped the 20 longest scaffolds assembled by ALLPATHS-LG (123 Mb in total) to our genome assembly (**Supplementary Fig. 5**) and found that 99% of these sequences were present in our assembly. Further, structural variations between the two assemblies only comprised ~0.9% of the mapped length.

We also assessed the assembly of genes. We sequenced mRNAs from four tissues (root, leaf, spica and stem), assembled the transcripts and mapped them to the genome assembly. Coverage of the gene regions was estimated to be ~96% (**Supplementary Table 7**). We then assessed coverage of core eukaryotic genes^{31,32} (**Supplementary Table 8**), which showed that >99% of conserved genes were present. We also mapped all the available gene sequences from foxtail millet present in the NCBI nucleotide database (28 genes) to the genome assembly, and of these, 27 genes were present in our draft genome (**Supplementary Table 9**). Finally, we mapped genes from *Arabidopsis*, sorghum, maize and rice to the genome to identify incomplete gene fragments (**Supplementary Table 10**). We randomly selected 82 gene fragments, and 81 (99%) of them were validated by Sanger sequencing (**Supplementary Table 11**).

Repeat content in the draft genome

Comprehensive repeat annotation of the draft Zhang gu genome revealed that ~46% of the genome comprised transposable elements (**Table 1** and **Supplementary Table 12**). Both retroelements (class I transposable element; 31.6%) and DNA transposons (class II transposable element; 9.4%) were identified. The most abundant repeat elements were long terminal repeats (LTRs), which comprised 29.6% of the genome. Of these LTRs, 22.1% of the genome was *gypsy* and 7.2% of the genome was *copia* (the rest 0.3% were other LTRs), with the ratio of *gypsy*-like to *copia*-like elements 3.1:1 (**Supplementary Table 13**). Other types of transposable elements were DNA transposons (9.4%), long interspersed nuclear elements (LINEs, 1.8%),

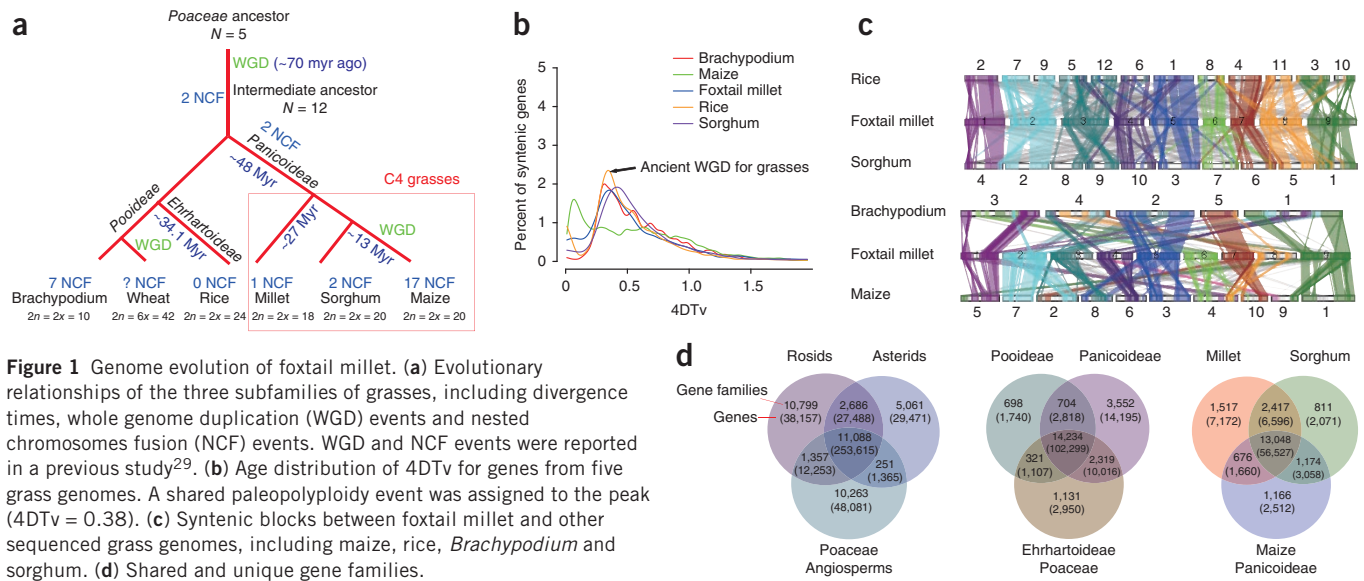


Figure 1 Genome evolution of foxtail millet. **(a)** Evolutionary relationships of the three subfamilies of grasses, including divergence times, whole genome duplication (WGD) events and nested chromosome fusion (NCF) events. WGD and NCF events were reported in a previous study²⁹. **(b)** Age distribution of 4DTV for genes from five grass genomes. A shared paleopolyploidy event was assigned to the peak (4DTV = 0.38). **(c)** Syntenic blocks between foxtail millet and other sequenced grass genomes, including maize, rice, *Brachypodium* and sorghum. **(d)** Shared and unique gene families.

short interspersed nuclear elements (SINEs, 0.2%) and other types (0.11%). There were 5.4% uncharacterized repeats. Transposable elements were more abundant in the low-recombination and gene-poor regions (**Supplementary Fig. 6**). The insertion time of most LTRs was estimated to be ~0.3–1.0 million years ago (Myr ago) (**Supplementary Fig. 7**). Recent LTRs (<0.1 Myr ago) comprised only 4.7% of the total LTRs. Distribution of recent LTRs was more random than that of ancient LTRs (**Supplementary Fig. 6**).

Transposable element proportions in the five grasses (*Brachypodium*³³, rice²⁸, sorghum²⁷, maize³⁴ and foxtail millet; **Supplementary Table 13**) varied from 27% (*Brachypodium*) to 84% (maize), with higher transposable element contents in larger genomes. In the case of LTRs, the proportions present in *Brachypodium*, rice (~20%) and foxtail millet (29.6%) were substantially lower than those present in sorghum (54%) and maize (75%). The composition of LTRs in these species varied, too. The ratio of *gypsy* LTRs to *copia* LTRs was highest in sorghum (3.7 to 1), lower in rice (2.83 to 1), *Brachypodium* (3.3 to 1) and foxtail millet (3.1 to 1), and lowest in maize (2.0 to 1). It has been reported that genome size changes that occur through rapid multiplication of transposable elements probably has large effects on specification^{35,36}. Therefore, different transposable element compositions in grasses resulting from transposable element multiplication might have affected the diversification of grasses.

Foxtail millet genes

We used an integrated annotation pipeline to predict a total of 38,801 genes in the draft genome (**Supplementary Table 14**). According to mRNA sequencing data from four tissues, ~81.7% of these genes were expressed. The average transcript length of annotated genes was 2,522 bp, which was substantially shorter than transcripts predicted by mapping the RNA sequencing data to the genome. To resolve this discrepancy, we added untranslated regions (UTRs) to the gene models, in accordance with the RNA sequencing data. In foxtail millet, the average intron length (442 bp), average exon length (256 bp) and average exons per gene (4.3) were similar to all other grasses except maize (**Supplementary Table 15**). Functional annotation confirmed that 78.8% of the genes have homologs with known functions in protein databases (**Supplementary Table 16**). We searched for conserved genes present in other grasses in the foxtail millet gene set,

and found that 99% (11,847 out of 11,969) of conserved genes have homologs in foxtail millet. This confirmed that the draft sequence is reasonably complete. We also identified 1,367 pseudogenes in the genome (**Supplementary Notes and Supplementary Table 17**).

Noncoding RNA genes were predicted in the genome (**Supplementary Table 18 and Supplementary Fig. 8**). We identified 99 ribosomal RNA genes, and four large clusters on chromosomes 8 and 9 harbored 23 rRNA genes (~23%). We identified 704 tRNA genes, compared to 611 in *Arabidopsis*³⁷. We observed large clusters of tRNA genes on chromosomes 1, 7, 8 and 9 (**Supplementary Fig. 8**). One hundred fifty-nine microRNA (miRNA) genes and 99 small nuclear RNA (snRNA) genes also showed biased chromosomal distribution (**Supplementary Fig. 8**).

Evolution of foxtail millet

The draft genome of foxtail millet was compared to *Brachypodium*, rice, sorghum and maize. The phylogenetic relationships among the grasses are shown in **Figure 1a**. Foxtail millet was split from sorghum and maize ~27 Myr ago. Paralogous relationships among the nine foxtail millet chromosomes revealed five major duplications between chromosomes 2 and 9, 4 and 1, 7 and 1, 6 and 2, and 5 and 3 (**Supplementary Fig. 9**). The accumulated transversion rate at fourfold degenerate sites of the third codons (4DTV) distribution of all paralogous genes further suggested that most of the duplications were generated in the whole genome duplication (WGD) event shared by all grasses (**Fig. 1b**), which was estimated to be ~70 Myr ago. Therefore WGD in foxtail millet preceded before the split from sorghum and maize.

Intergenomic analyses between foxtail millet, *Brachypodium*, rice, sorghum and maize revealed highly conserved colinearity, which supports a close evolutionary relationship among these grasses. In total, we identified 24 large colinear blocks between foxtail millet and rice; 19 between foxtail millet and sorghum; 20 between foxtail millet and *Brachypodium*; and 29 between foxtail millet and maize, indicating that 71.8%, 72.1%, 61.5% and 86.7% of the foxtail millet genome is colinear with these grasses, respectively (**Fig. 1c and Supplementary Table 19**).

Evolution of chromosomes

Despite the close evolutionary relationship among the grasses, the chromosomes of grasses have undergone extensive rearrangements.

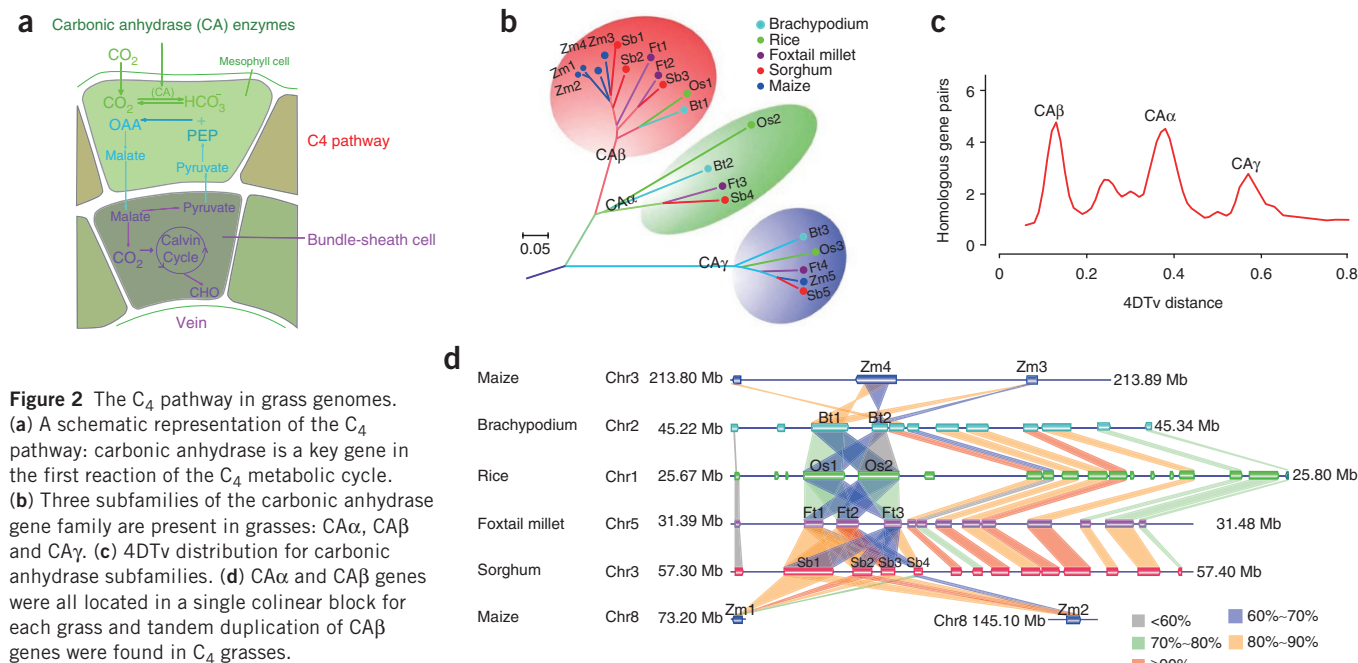


Figure 2 The C_4 pathway in grass genomes. (a) A schematic representation of the C_4 pathway: carbonic anhydrase is a key gene in the first reaction of the C_4 metabolic cycle. (b) Three subfamilies of the carbonic anhydrase gene family are present in grasses: $CA\alpha$, $CA\beta$ and $CA\gamma$. (c) 4DTV distribution for carbonic anhydrase subfamilies. (d) $CA\alpha$ and $CA\beta$ genes were all located in a single colinear block for each grass and tandem duplication of $CA\beta$ genes were found in C_4 grasses.

Ancestral chromosome reconstruction of grasses has revealed that the ancestor of grasses had 12 chromosomes after one WGD and two nest chromosome fission events³⁸. However, how different species evolved from the 12 ancestral chromosomes is unknown. To investigate evolution of foxtail millet chromosomes, we used rice as the reference comparison strain because it has retained 12 chromosomes. We found that foxtail millet chromosomes 2, 3 and 9 were colinear with rice chromosomes 7 and 9, 5 and 12, and 3 and 10, respectively (Fig. 1c), indicating that after divergence from the common ancestor, three pairs of chromosomes separately fused to form three chromosomes in foxtail millet. Among the three nest chromosome fusion events that occurred in foxtail millet, two also occurred during the evolution of sorghum (Fig. 1c), indicating that these two nest chromosome fusion events most likely occurred before the divergence of sorghum and foxtail millet. The nest chromosome fusion event that fused chromosomes 8 and 9 of sorghum into chromosome 3 of foxtail millet likely occurred after divergence of foxtail millet from sorghum. Such chromosome reshuffling events might have introduced genetic variation and contributed to speciation of these plants.

Evolution of gene families

Differences in gene families of grass genomes are important sources of genetic traits and adaptation in different species^{39,40}. Since divergence from their common ancestor, grass genomes have an active history of rearrangements that have been accompanied by both minor- or macro-colinearity erosion, which could have a significant impact on gene inventories. In foxtail millet, many genes that underwent ancestral duplication in the grass paleopolyploidy event have since lost one copy. We identified 26,369 (69.5%), 24,722 (65.2%), 25,717 (67.8%), 12,167 (32.1%) foxtail millet genes that were in colinear blocks with rice, sorghum, *Brachypodium* and maize genes, respectively. Genes in colinear blocks should have paralogs that arose during the WGD. However, within those colinear blocks, 37.1% of rice, 34.7% of sorghum, 34.8% of *Brachypodium* and 29.9% of maize genes retained only one copy. This indicates that nearly half of the duplicated gene pairs lost one copy over time. The other half of the

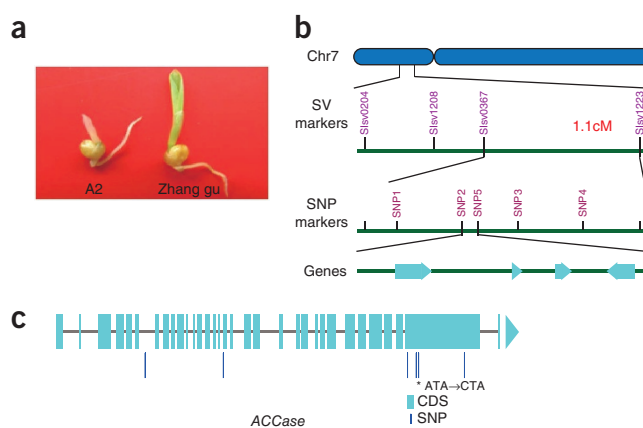
duplicated genes retained both copies, and would be expected to be subjected to neofunctionalization or subfunctionalization.

Expansion or contraction of gene families has an important role in diversification of flowering plants^{41,42}. We identified gene families in fifteen plants, including plants from the rosids clade (*Arabidopsis*, papaya, populus, cucumber, soybean, apple, grape, castor bean), asterids clade (potato) and poaceae family (foxtail millet, rice, maize, sorghum, *Brachypodium*) (Fig. 1d). In total, we identified 15,382 plant gene families shared by all fifteen species and 11,088 pan-grass gene families that are shared by all angiosperm plants, most of which were involved in molecular binding, transport, catalytic activity and metabolic processes. Comparing species from poaceae with species from other grass clades, we identified 10,263 gene families specific to poaceae. Additional comparisons within poaceae species revealed 3,552 panicoideae-specific families and 1,517 foxtail millet-specific gene families. Gene Ontology annotation of these foxtail millet-specific gene families showed that some of these gene families may have roles in stress responses (Supplementary Table 20). One such family that contains 586 genes was functionally annotated as “response to water.” This indicates that foxtail millet-specific genes might be related to adaptation of foxtail millet to semi-arid environments. In addition to common or specific gene families, we also identified 591 gene families in foxtail millet that showed considerable contraction in foxtail millet compared to other grasses (Supplementary Table 21).

C_4 pathway in grasses

C_4 plants have an advantage over C_3 plants because C_4 photosynthesis has more efficient carbon fixation, especially in arid areas^{18,43}. C_4 photosynthesis is of polyphyletic origin, which indicates that relatively few evolutionary changes are required to establish this photosynthetic system⁴⁴. We identified the key photosynthetic genes involved in the C_4 metabolic cycle (Fig. 2a) in *Brachypodium*, rice, sorghum, maize and foxtail millet using known C_4 pathway genes. We found that all the genes involved in C_4 carbon fixation pathways were also present in C_3 plants (Supplementary Table 22). Thus the evolution of the C_4 pathway may have resulted from functional changes in these genes.

Figure 3 Mapping sethoxydim resistance. (a) Sethoxydim resistance of Zhang gu compared with the sethoxydim sensitive phenotype of A2. (b) Mapping of the trait by structural variant markers as well as fine mapping using SNPs. (c) Variations located within the acetyl-CoA gene (Millet_GLEAN_10024326) between Zhang gu and A2. The fine mapping of the sethoxydim trait was consistent with previous studies of this trait in green millet¹⁵, in which experimental validation of the gene function was also carried out.



We also analyzed the carbonic anhydrase genes, which have important roles in the CO₂-concentrating process of C₄ plants. A phylogenetic tree of the carbonic anhydrase homologs identified in five grasses revealed that there were three major carbonic anhydrase subfamilies (CA α , CA β and CA γ), consistent with previous data^{14,45} (Fig. 2b). There was one copy of CA α and CA γ in each of the five species, but there were more copies of CA β in the C₄ plants than in rice and *Brachypodium*. Distribution of 4DTv of all the identified carbonic anhydrase homolog genes (Fig. 2c) suggested that CA γ genes are more ancient than genes from the other two subfamilies and CA α genes were probably generated in the WGD of grasses (4DTv of CA α was 0.38, coincident with the 4DTv of the WGD). CA β genes were likely to be generated by tandem duplication (Fig. 2d). We identified one CA α gene, two CA β genes and one CA γ in foxtail millet. The CA α and CA β genes formed a gene cluster on chromosome 5 (~27 Mb) in a colinear block with the homologous counterparts from *Brachypodium* (chromosome 2, ~45.25 Mb), rice (chromosome 1, ~26.5 Mb), sorghum (chromosome 3, ~57.3 Mb) and maize (chromosomes 3 and 8). Of the three carbonic anhydrase subfamilies, CA β catalyzes the first reaction in the C₄ pathway by hydrating atmospheric CO₂ to bicarbonate in the cytosol of mesophyll cells. Tandem duplication of CA β genes was only observed in C₄ grasses, so it may be important in the evolution of the C₄ pathway.

Combining this phylogenetic analysis with the transcriptome data generated in this study, we found that one of the foxtail millet CA β genes, *Ft_CA1* (Millet_GLEAN_10030850), was highly expressed in mesophyll (with fragments per kilobase of exon per million fragments mapped, FPKM value 22,970). The nonsynonymous to synonymous ratio of this gene (9.9) revealed that it was under adaptive selection (Supplementary Fig. 10). These findings indicate that this *Ft_CA1* gene might be important for the C₄ photosynthesis pathway in foxtail millet.

Exploitation of the foxtail millet genome

Availability of a genome sequence facilitates quantitative trait mapping. We used the foxtail millet genome to map herbicide resistance. The Zhang gu strain is resistant to sethoxydim whereas the A2 strain is sensitive to sethoxydim (Fig. 3a). In the F₂ population of Zhang gu and A2 that was used for genetic map construction, the segregation of resistance to sethoxydim compared with sensitivity to sethoxydim was about 3:1 (Supplementary Table 23), which suggested that sethoxydim resistance was probably related to a single, completely dominant, nuclear gene. We were able to map the trait to one 1.1-cM (2.4-Mb) region between marker SIsv0367 and marker SIsv1223 using our genetic map (Fig. 3b). To fine-map this trait, we chose five SNP markers (Supplementary Table 24) located between SIsv0367 and SIsv1223, and genotyped them in the F₂ population. The trait of sethoxydim resistance was located between SNP markers 2 and 5 (Fig. 3b), in a ~100-kb region. Four genes were located in that region, including *Ft_SRI*, Millet_GLEAN_10024326, which was annotated as acetyl-CoA carboxylase. Acetyl-CoA carboxylase is the target enzyme of sethoxydim, and mutations in this gene would result in resistance to sethoxydim¹⁵. Within this gene, there were only six SNPs (Fig. 3c). Of those SNPs, one possible functional

SNP was located in the coding sequence of this gene and caused a codon change from CTA (Zhang gu) to ATA (A2) which changed leucine (Leu, in Zhang gu) to isoleucine (Ile, in A2). In green millet, the substitution from Ile to Leu in one acetyl-CoA gene has already been shown to code for sethoxydim resistance^{15,46}. Therefore, we have shown that the draft genome sequence can be used to accurately identify the gene for sethoxydim resistance.

DISCUSSION

The genetic and genomic data of economically important grasses are an invaluable resource for comparative and functional genomics studies. We have added to the resource of grass genomes with the draft genome sequence of foxtail millet. To assemble the foxtail millet genome we applied a whole genome shotgun strategy combined with next-generation sequencing technology. Foxtail millet has a small genome (~490 M), which is rich in repeats, consistent with the sequenced genomes of other grass species, and low in heterozygosity. We assembled the genome using data from next-generation sequencers and mapped the sequences onto chromosomes using a genetic map. The genome assembly covers ~86% of the estimated genome size, and the unassembled part is largely composed of repeat elements. Repeat-rich regions of the genome, especially long repeats that cannot be resolved using large insert-size libraries, are difficult to assemble, which resulted in a large number of contigs and scaffolds. However, for regions lacking repeat sequences, the assembly was easier and most of the genome was present in just a few hundred long scaffolds. This draft genome will be improved if bacterial artificial chromosome sequencing data of foxtail millet is developed.

The draft genome of foxtail millet will provide an important resource for crop improvement. The provision of a set of genetic markers through resequencing of a second foxtail millet strain (A2) enabled us to develop a genetic map. We also generated a set of RNA data from different tissues to underpin future studies of gene function. Finally, we applied the genome sequence to map an herbicide-resistance trait, to demonstrate the utility of this draft genome in crop improvement. Combining these genome and transcriptome data with a germplasm collection, gene transformation method and other available resources, foxtail millet can now be studied and improved in an effective way.

METHODS

Methods and any associated references are available in the online version of the paper.

Data access The genomic reads of Zhang gu and A2, as well as RNA sequencing data of Zhang gu, have been deposited into NCBI Short

Read Archive (SRA) under accession number SRA048234. The information for the raw reads data can be found in **Supplementary Table 25**. The genome sequence and annotation data set have been deposited into NCBI (Project ID: PRJNA77795 (chromosomes) and PRJNA73995 (scaffolds), the accession number is GSM892310). The genome assembly, annotation as well as other data mentioned can be downloaded from ftp://ftp.genomics.org.cn/pub/Foxtail_millet. Genome browser can be found on the foxtail millet genome website (<http://foxtailmillet.genomics.org.cn/>). The primer sequence of 759 markers used to construct the genetic map can be found in **Supplementary Table 3**. The phenotype data of 480 F₂ individuals are in **Supplementary Table 24**. The additional genetic markers can be found in **Supplementary Table 25**.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank L. Goodman for editing the manuscript and H.-M. Lam for providing valuable comments. This work was supported by the following funding to BGI-Shenzhen: Shenzhen-National Science and Technology Plan Projects (ZYC200903240077A), Basic Research Program of Shenzhen (JC2010526019), Key Laboratory Project Supported by Shenzhen City (CXB200903110066A), Shenzhen Hundred-million Project, Chinese 948 Program supported by the Ministry of Agriculture (2010-Z31), National Science and Technology Ministry Project -973 program (2007CB815703), National Science and Technology Ministry Project -973 program (2007CB815705), National Natural Science Funds for Distinguished Young Scholar (30725008).

AUTHOR CONTRIBUTIONS

Jun W., G.Z., Z.Z. and Jian W. designed the project. Jun W., Jian W., G.Z., X.L., X.X., Junyi W. and Y.L. led the sequencing and analysis. Z.Y., S.P., Q.X. and X.P. did the genome assembly. M.X., X.Y. and D.Z. did the annotation. P.Z., W.W., C.B. and R.C. did the evolutionary analysis. Z.Q., J.H., Y.Z., Henan L., Hua L., N.L. and C.W. constructed the genetic map. Y.T. and C.H. did the optical mapping analysis. Z.Q., R.W., T.G., Y.C., P.H., C.L., H.X., Q.S. and Q.C. conducted the trait mapping analysis. X.L., S.C., Z.Q., X.X. and G.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/doi/10.1038/nbt.2195>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA) license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>.

- Zohary, D. & Hopf, M. Domestication of Plants in the Old World: the Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley. Third edn. (Oxford University Press, 2000).
- Doust, A.N., Kellogg, E.A., Devos, K.M. & Bennetzen, J.L. Foxtail millet: a sequence-driven grass model system. *Plant Physiol.* **149**, 137–141 (2009).
- Wang, Z.M., Devos, K.M., Liu, C.J., Wang, R.Q. & Gale, M.D. Construction of RFLP-based maps of foxtail millet, *Setaria italica* (L.) P. Beauv. *Theor. Appl. Genet.* **96**, 31–36 (1998).
- Devos, K.M., Wang, Z.M., Beales, J., Sasaki, T. & Gale, M.D. Comparative genetic maps of foxtail millet (*Setaria italica*) and rice (*Oryza sativa*). *Theor. Appl. Genet.* **96**, 63–68 (1998).
- Devos, K.M., Pittaway, T.S., Reynolds, A. & Gale, M.D. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. *Theor. Appl. Genet.* **100**, 190–198 (2000).
- Wang, C. *et al.* Population genetics of foxtail millet and its wild ancestor. *BMC Genet.* **11**, 90 (2010).
- Fukunaga, K., Wang, Z., Kato, K. & Kawase, M. Geographical variation of nuclear genome RFLPs and genetic differentiation in foxtail millet, *Setaria italica* (L.) P. Beauv. *Genet. Resour. Crop Evol.* **49**, 95–101 (2002).
- Schontz, D. & Rether, B. Genetic variability in foxtail millet, *Setaria italica* (L.) P. Beauv.: Identification and classification of lines with RAPD markers. *Plant Breed.* **118**, 190–192 (1999).
- Fukunaga, K.F., Kawase, M.K. & Kato, K.K. Structural variation in the Waxy gene and differentiation in foxtail millet [*Setaria italica* (L.) P. Beauv.]: implications for multiple origins of the waxy phenotype. *Mol. Genet. Genomics* **268**, 214–222 (2002).
- Jia, X.-P. Development of EST-SSR in foxtail millet (*Setaria italica*). *Genet. Resour. Crop Evol.* **54**, 233–236 (2007).
- Zhang, J. *et al.* Construction and application of EST library from *Setaria italica* in response to dehydration stress. *Genomics* **90**, 121–131 (2007).
- Langdale, J.A. C4 cycles: past, present, and future research on C4 photosynthesis. *Plant Cell* **23**, 3879–3892 (2011).
- Monson, R.K. Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis. *Int. J. Plant Sci.* **164**, S43–S54 (2003).
- Ludwig, M. Carbonic anhydrase and the molecular evolution of C4 photosynthesis. *Plant Cell Environ.* **35**, 22–37 (2012).
- Délye, C., Wang, T. & Darmency, H. An isoleucine-leucine substitution in chloroplastic acetyl-CoA carboxylase from green foxtail (*Setaria viridis* L. Beauv.) is responsible for resistance to the cyclohexanedione herbicide setoxydim. *Planta* **214**, 421–427 (2002).
- Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
- Kim, E.B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223–227 (2011).
- Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29**, 735–741 (2011).
- Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
- Wang, X. *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
- Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
- Alkan, C., Sajadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
- Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
- Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
- Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
- Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38**, 95–109 (2011).
- Paterson, A.H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
- International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
- Murat, F. *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557 (2010).
- Zhou, S. *et al.* A single molecule scaffold for the maize genome. *PLoS Genet.* **5**, e1000711 (2009).
- Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
- Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
- The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
- Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
- Bennetzen, J.L. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269 (2000).
- Hu, T.T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24 (2008).
- Velasco, R. *et al.* The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
- Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011).
- Purugganan, M.D., Rounsley, S.D., Schmidt, R.J. & Yanofsky, M.F. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* **140**, 345–356 (1995).
- Lashbrook, C.C., Tieman, D.M. & Klee, H.J. Differential regulation of the tomato ETR gene family throughout plant development. *Plant J.* **15**, 243–252 (1998).
- Sage, R.F. & Monson, R.K. *C4 Plant Biology* (Academic Press, 1999).
- Westhoff, P. & Gowik, U. Evolution of C4 photosynthesis—looking for the master switch. *Plant Physiol.* **154**, 598–601 (2010).
- Tiwari, A., Kumar, P., Singh, S. & Ansari, S. Carbonic anhydrase in relation to higher plants. *Photosynthetica* **43**, 1–11 (2005).
- Wang, T. & Darmency, H. Inheritance of setoxydim resistance in foxtail millet, *Setaria italica* (L.) Beauv. *Euphytica* **94**, 69–73 (1997).

ONLINE METHODS

Sample preparation. The widely cultivated commercial line Zhang gu was collected from northern China. The other commercial line A2 was collected from Zhangjiakou, China. The Zhang gu line was herbicide (sethoxydim) resistant. A hybrid population of 480 individuals was bred from these two lines. The F₂ population was used in construction of a genetic map and QTL analysis. Genomic DNA was isolated from young leaves using a modified SDS method⁴⁷.

Library construction. Paired-end libraries with 170 bp, 200 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb, respectively, were constructed following a standard protocol provided by Illumina. We used a fosmid library construction method to construct the 40 kb library (see details in **Supplementary Notes**, “Construction of 40 kb insert size library,” **Supplementary Fig. 11**).

Genome sequencing. The sequencing was performed using Illumina Genome Analyzer II and HiSeq 2000 (short insert size libraries were sequenced by GA II and long insert size libraries, 5 k, 10 k, 20 k, 40 k, were sequenced by HiSeq 2000, just for convenience of sequencing and data output) according to the manufacturer's standard protocol. Low-quality reads, reads with adaptor sequences and duplicated reads were filtered, and the remaining high-quality data was used in the assembly.

Genome assembly. Genome size was estimated by k-mer analysis (see details in **Supplementary Notes**, “Estimation of the genome size,” **Supplementary Fig. 12**). SOAPdenovo²³ was applied to assemble the genome in the procedure of contig construction, scaffold construction and gap closure. Then a genetic map was constructed for anchoring the scaffolds onto chromosomes. We developed genetic markers by resequencing strain A2. We constructed a short insert size library (500 bp) of A2 from genomic DNA of young leaves and Illumina GA II was used to do the sequencing. In total ~5G raw reads were generated. After mapping these reads back to the scaffolds of the Zhang gu strain, variations were identified (**Supplementary Fig. 13**, **Supplementary Fig. 14** and **Supplementary Fig. 15**). SOAPsnp⁴⁸ (Version 1.02) was used to identify SNPs. SOAPindel (<http://soap.genomics.org.cn/>, Version 1.01) was applied to identify indels. SOAPsv (<http://soap.genomics.org.cn/>, Version 1.01) was used to identify structure variations.

In order to construct the genetic map, we selected 641 structural variants and 118 SNPs in the 613 longest scaffolds. For structural variants, in order to facilitate the genotyping, we selected structural variants with a size change of ~500 bp, which were determined by PCR and agarose gel electrophoresis. Primers were then designed based on the genome sequences of both structural variant and SNP markers. PCR was used to amplify the corresponding region in 480 F₂ individuals generated by crossing Zhang gu and A2. Genotypes of structural variants were then determined by agarose gel electrophoresis and genotypes of SNPs were determined by Sanger sequencing. We constructed the genetic map using JoinMap⁴⁹. The genetic map was used to anchor the 631 longest scaffolds onto 9 chromosomes.

Assessing the genome assembly. To assess the genome assembly, we first mapped all the reads to the assembled genome by SOAP2 (ref. 48). Only mapped reads were then used to assess the support of assembly in regions. Then we compared the assembled genome sequence to the genome of sorghum²⁷ by Mummer⁵⁰ (Version 3.0) and colinear regions were thus identified. Rearrangements were then assessed by the support of paired-end mapped reads.

Optical mapping was then conducted. Large amounts of DNA from the Zhang gu strain was extracted from the young leaves for construction of optical maps. Whole genome shotgun single-molecule restriction maps were then generated with the highly automated Argus system (OpGen) based on optical mapping technology. Overall, 252,267 single-molecule restriction maps (SMRMs) (>250 kb) with an average size of 271 kb were generated. The total size was about 68.9 Gb, ~140× depth of the genome. The optical data were mapped to the genome assembly using commercial software (OpGen). Briefly, the fragments of optical data were mapped back to the genome according to the restriction enzyme recognition sites and the length of the fragments. The mapping accuracy of the restriction enzyme recognition sites was then determined by that software.

ALLPATHS-LG²⁴ (Release 37250) was then applied to assemble the genome. The assembled large scaffolds by ALLPATHS-LG were compared to our genome assembly by the software, Mummer.

RNA sequencing. Four tissues (root, stem, leaf and spike) from young seedlings were used to isolate RNA. Total RNA was extracted using Trizol (Invitrogen). The isolated RNA was then treated by RNase-Free DNase, and then subsequently treated using Illumina mRNA-Seq Prep Kit following the manufacturer's instruction. The insert size of the RNA libraries was about 200 bp, and the sequencing was done using Illumina GA II. Raw reads were filtered if there were adaptor contaminations and low quality (>10% bases with unknown quality). After filtering, we assembled the RNA data by Trinity⁵¹ (r2011-08-20) of four tissues separately.

Assessing the gene region coverage. We assessed the gene region coverage of the genome assembly by using the assembled RNA sequences, core eukaryotic genes, publicly available foxtail gene sequences and incomplete gene fragments in the genome. RNA assembly of four tissues was mapped back to the genome assembly using BLAT⁵². The coverage was then calculated based on the mapping result. Core eukaryotic genes identified by CEGMA³² and the foxtail millet gene sequences archived in NCBI nucleotide database (28 genes) were also mapped back to the genome assembly by BLAT to calculate the gene region coverage. Proteins of *Brachypodium*, rice, maize and sorghum were mapped to foxtail millet genome by tblastn. Gene fragments were thus identified. And we filtered redundant gene fragments and retained fragments longer than 150 bp. The fragments were then subjected to a *de novo* gene annotation by AUGUSTUS⁵³ to identify the incorrect start codons, premature stop codons and other incorrect codons. In total we identified ~13,000 gene fragments, and we randomly selected 81 of them to be validated by PCR and Sanger sequencing.

Repeat annotation. Transposable elements were identified in the genome using a combination of *de novo* and homology-based approaches. First, we used three *de novo* software programs (LTR_FINDER⁵⁴ (Version 1.0.3), PILER⁵⁵ and RepeatScout⁵⁶ (Version 1.05)) to build *de novo* repeat database of foxtail millet. We then used RepeatMasker⁵⁷ (Version 3.2.7) to identify repeats using both the repeat database we had built and Repbase⁵⁸. We also used RepeatProteinMask (<http://www.repeatmasker.org/>, Version 3.2.2) to search the protein database in Repbase against the genome to identify repeat related proteins. We annotated the tandem repeats in the foxtail millet genome using RepeatMasker and TRF⁵⁹. Then we combined the *de novo* prediction, the homolog prediction of transposable elements according to the coordination in the genome. Complete LTRs were predicted by LTR_STRUC⁶⁰. Muscle was used to align LTRs and divergence was calculated using distmat in EMBOSS package. Divergent time was estimated by setting *Brachypodium* as reference. The divergent time of LTRs in *Brachypodium*³³ was used.

Several steps were followed to classify transposable elements: (i) blastn against Repbase, (ii) blastx against transposable element proteins, (iii) blastn against plant/animal repeat databases, (iv) blastx against SwissProt⁶¹ proteins, (v) tblastx against Repbase and (vi) tblastx against plant/animal repeat databases. For each step, the transposable elements having confident hits with known repeats were classified either at the DNA level (E-value ≤ 10⁻¹⁰, identity ≥ 80%, coverage ≥ 30% and the minimal matching length ≥ 80 bp) or at the protein level (E-value ≤ 10⁻⁴, identity ≥ 30%, coverage ≥ 30% and the minimal matching length ≥ 30 amino acid). LTR retrotransposons identified by LTR_FINDER were classified as “unclassified LTR” if no homology to known repeats was found.

To infer the insertion time of LTRs, full-length LTRs were identified using LTR_STRUC⁶⁰ with default parameters. The insertion time was estimated based on the sequence divergence between two LTRs of the same element. All the LTRs were aligned by MUSCLE⁶², and then the distance between them was calculated by the Kimura two-parameter model using distmat program implemented in the EMBOSS package (<http://emboss.sourceforge.net/>). Then the same procedure was done to the *Brachypodium* genome, and according to the insertion time of LTRs in *Brachypodium* reported³³, the insertion time of LTRs in the foxtail millet genome was estimated.

Gene prediction. We conducted the gene annotation by the following steps: (i) *De novo* gene prediction. We performed *de novo* predictions on repeat

masked genome assembly. We used AUGUSTUS⁵³ (Version 2.03) and Fgenesh⁶³ (Version 1.3) to do the *de novo* annotation. (ii) Homolog prediction. We mapped the protein sequences from *Arabidopsis*, rice, sorghum and maize to the foxtail millet genome using tblastn, by an E-value cutoff of 10^{-5} , and then Genewise⁶⁴ (Version 2.2.0) was used for gene annotation. (iii) RNA-aided annotation. We mapped all the RNA reads back to the reference genome by Tophat⁶⁵ (Version 1.0.14, implemented with bowtie⁶⁶ Version 0.12.5) and assembled the transcripts according to the genome using Cufflinks⁶⁷ (Version 0.8.2). Open reading frames (ORFs) were predicted using BESTORF (<http://linux1.softberry.com/berry.phtml?topic=bestorf&group=programs&subgroup=gfind>) with parameters trained on monocot genes without filtering of UTRs. (iv) GLEAN⁶⁸. All the predictions were combined using GLEAN to produce the consensus gene sets.

Gene functions were assigned according to the best match of the alignments using blastp (E-value $\leq 10^{-5}$) searching against SwissProt and Uniprot databases⁶¹ (Release 15.10). The motifs and domains of genes were determined by InterProScan⁶⁹ (Version 4.5) against protein databases. Gene Ontology⁷⁰ IDs for each gene were obtained from the corresponding InterPro entry. All genes were aligned against KEGG⁷¹ (Release 48.2) proteins, and the pathway in which the gene might be involved was derived. If the best hit of the genes in any of these process was “function unknown”, “putative”, the second-best hits were used to assign function until there were no more hits that meet the alignment criteria (then this gene is determined as functionally unknown).

The tRNA genes were identified by tRNAscan-SE⁷². For rRNA identification, we first downloaded the *Arabidopsis* rRNA sequences from NCBI (<http://www.ncbi.nlm.nih.gov/guide/dna-rna/>). Then rRNAs in the database were aligned against the foxtail millet genome using blastn to identify possible rRNAs. Other noncoding RNAs, including miRNA, snRNA, were identified using INFERNAL⁷³ by searching against the Rfam database.

Evolution analysis. We first identified homologous gene families across all the available grass genomes (*Brachypodium*, rice, maize, sorghum and foxtail millet) using OrthoMCL⁷⁴. Gene families for which there is a single gene in each species (single-copy gene families) were then extracted. For each of these single-copy gene families, proteins of the homolog genes were aligned by muscle⁷⁵ and the alignments were CDS-back-translated into CDS using an in-house perl script. Then all the back translated CDS sequences were joined into one super gene for each species. Fourfold generation sites were then identified in the super genes of all five species and the transversion rates (4DTv) between every two species were then calculated by an in-house perl script. Then using the 4DTv to represent distances between every two species we applied MrBayes⁷⁶ (Version 3.2.0) to construct a phylogenetic tree. Finally MultiDivergence⁷⁷ was applied to infer the divergence time base on the phylogenetic tree constructed (setting the divergence time of rice and *Brachypodium* to 34 Myr ago in previous study³³). The topology was simplified into a sketch map in **Figure 1** with wheat added. The estimated divergent time was assigned onto branches.

To identify syntenic blocks within foxtail millet genome, we used MCScan⁷⁸, which set genes as anchors. First we used blastp to align foxtail millet protein data set to itself (setting E-value to be 10^{-7}). The alignment result was then subjected to MCScan to determine syntenic block. For gene pairs in the syntenic blocks identified by MCScan, we calculated the 4DTv of each gene pair and plotted the distribution of the 4DTv of all the gene pairs.

Identification of chromosome reshuffling. After identification of the syntenic and colinear blocks between foxtail millet with other grasses, we set rice genome as reference and identified all the homologous chromosome relationships between rice and foxtail millet, rice and sorghum, foxtail millet and sorghum, foxtail millet and maize.

Gene family analysis. We clustered paralogs and orthologs using OrthoMCL method⁷⁴ with the default settings (blastp E-value $\leq 10^{-5}$) for 15 sequenced plants, including eudicots, asteroids and grasses. After getting all gene families, we classified the families according to the presence or absence of genes for specific species and determined which gene families were species-specific or genus-specific.

Identification of genes in C₄ photosynthesis pathway. We retrieved the photosynthesis pathway genes from two resources: (i) from the NCBI database

(we downloaded the genes identified to be involved in photosynthesis of sorghum, maize and rice); and (ii) from the functional annotation of foxtail millet proteins.

Trait mapping of sethoxydim. We used the F₂ populations (480 individuals) generated by crossing between the Zhang gu (sethoxydim resistant) and A2 (sethoxydim sensitive) strains. Sethoxydim-resistant phenotype data were collected in the F₂ population by seeding on sethoxydim medium (50 mg/l). The seeds of F₂ plants were grown in the sethoxydim medium to determine the phenotypes. If seeds were resistant to sethoxydim, the phenotype of the F₂ individual was resistant to sethoxydim. If the seeds were segregated in terms of resistance to sethoxydim, the phenotype of the F₂ individual was heterozygous.

All the in-house scripts used are available on request.

47. Peng, J., Wang, H., Haley, S.D., Peairs, F.B. & Lapitan, N.L.V. Molecular mapping of the Russian wheat aphid resistance gene in wheat. *Crop Sci.* **47**, 2418–2429 (2007).
48. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
49. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744 (1993).
50. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
51. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
52. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
53. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–312 (2004).
54. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
55. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–i158 (2005).
56. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005).
57. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Prot. Bioinform.* **25**, 4.10.1–4.10.14 (2009).
58. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
59. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
60. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362–367 (2003).
61. The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**, D169–D174 (2009).
62. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
63. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
64. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
65. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
66. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
67. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
68. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
69. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
70. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
71. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
72. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
73. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
74. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
75. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
76. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
77. Thorne, J.L. & Kishino, H. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**, 689–702 (2002).
78. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).