Supplementary Information

This file includes:

Nature Biotechnology: doi:10.1038/nbt.2195

Nature Biotechnology: doi:10.1038/nbt.2195

Supplementary Notes

Construction of 40 kb insert size library

For the short insert size libraries (170~800 bp), 5 μg DNA were used to be sheared to fragments with size of 170~800 bp, end paired, A-tailed and ligated to sequencing adapters. Then those fragments were selected according to sizes (170 bp, 200 bp, 500 bp and 800 bp) on the agarose gel and amplified by LM-PCR to yield the short insert size libraries. For long insert size libraries, 20~40 μg DNA were sheared to corresponding size using nebulization for 2 kb and HydroShear (Covaris, Woburn, MA) for 5 kb, 10 kb and 20 kb. Next, the DNA fragments were end-repaired using biotinylated nucleotide analogues, size selected at 2, 5, 10 and 20 kb circularized by intramolecular ligation. Circular DNA molecules were sheared with Adaptive Focused Acoustic (Covaris) to an average size of 500 bp. Biotinylated fragments were purified on magnetic beads (Invitrogen, CA, USA), end-repaired, A-tailed and ligated to Illumina paired-end adapters, size-selected again and purified by LM-PCR.

For 40 kb insert size library, we applied a fosmid library construction by the following steps (as in **Supplementary Fig. 11**).

Step 1 Construction of fosmids

More than 100 μg DNA were extracted and sheared gently to large size fragments. DNA fragments were separated by pulsed field gel electrophoresis, and 40 kb fragments were selected. Those selected fragments were ligated to fosmid vector (CopyControl™ Fosmid Library Production Kit, Epicentre, USA) following production manual.

Step 2 Construct the circular DNA with vector and ends of 40 kb DNA fragments

Fosmid plasmids were then fragmented in order to get constructs which contain the ends from a large fragment of DNA. Approximately 20 μg pooled plasmid DNA was sheared (GeneMachine, San Carlos, CA., USA, 20cycles, code 8) and purified with QIAquick PCR Purification Kit (Qiagen, Germantown, MD),

Nature Biotechnology: doi:10.1038/nbt.2195

and then it was end-repaired to generate blunt, 5'-phosphorylated ends. End-repaired DNA was then resolved by 0.6% Megebase agarose gel electrophoresis (5V/CM, 16 hours).

DNA fragments migrating to 8.4-9.0kb in length were excised and purified by QIAquick Gel Purification Kit. In order to form a second circular DNA we proceeded with the ligation of the DNA fragment from the previous step by T4 DNA ligase (Invitrogen, CA, USA) to form a second circular DNA (16℃,16h), digested linear DNA(16 μl 100 mM ATP,192 μl 10 x Plasmid-Safe ATP-dependent DNase buffer,48μl Exonuclease I), the reaction solution was incubated at 37℃ for 30minutes,and then 72℃ for 20 minutes. 64 μl 0.5M EDTA was added to the solution and a process of purification followed (QIAquick PCR Purification Kit). Linear DNA comprised by two end regions of a large fragment of DNA were then generated by PCR (5 μl 10 x Ex *Taq* buffer,4 μl 2.5mM dNTP,2ul 10um primer and 0.25 μl Ex *Taq*). The primer sequences were PCC2F : 5'-CAGGAAACAGCCTAGGAA-3' and PCC2R: 5'-GTACAACGACACCTAGAC-3'.Then, adaptors were added to the linear DNA after purification (Qiagen MinElute PCR Purification Kit). In order to enrich the purified (Qiagen MinElute PCR Purification Kit) DNA for Illumina sequencing, 5μl 10 x Ex *Taq* buffer, 4 μl 2.5 mM dNTP, 0.25 μl Ex *Taq* and 10 μM sequencing primer (forward sequencing primer was PE1.0 and the reverse sequencing primer sequence was PE2.0) were added to the DNA to perform PCR. PCR production was denatured by heating (65℃, 10min and then placed them on the ice) and then separated by 2.0% low range agarose gel electrophoresis (15V/CM, 2 hours). After staining, fragments from 400-700bp were excised from the gel and purified (Qiagen MinElute Gel Purification Kit) before sequencing.

Estimation of the genome size

Genome size would be related with the total length of sequenced reads and the sequencing depth, thus can be derived from the k-mer distribution[1]. k-mer refers to a

Nature Biotechnology: doi:10.1038/nbt.2195

sequence with the length of k bp, and each unique k-mer within a genome dataset can be used to determine the discrete probability distributions of all possible k-mers and their frequency of occurrence as in previous studies[1,2].

The estimation of genome size by k-mer frequency distribution analysis is sensitive to sequencing errors, so we just used the filtered reads from the short insert size libraries of which the duplication rate is low. The sequencing depth was estimated by determining the peak value of the frequency curve of k-mer occurrence distribution. Genome size was then estimated by modified Lander-Waterman algorithm[3], G= (N×(L-K+1)-B)/D, where N is the total number of reads, L is the average length of reads, and K is k-mer length (set to 17 in our analysis). To minimize the influence of sequencing errors, k-mers with low frequency (B), occurring less than 4 times, were discarded. G is the genome size, and D is the overall depth estimated from k-mer distribution (pkdepth). In the k-mer analysis we conducted, the read number is 115,144,638, and k-mer size is 17, thus the number of k-mer is 6,793,533,642. By plotting the occurrence of k-mers against the percentage of the corresponding k-mers (**Supplementary Fig. 12**), we defined the peak depth to be 14. Thus the genome size was estimated to be 485 Mb by the k-mer analysis.

Identification of pseudogenes

To detect pseudogenes, we applied a pipeline described in previous study[4]. We first mapped all the expressed proteins (which is supported by uniquely mapped RNAseq reads, ie, FPKM value >0) to the millet genome assembly using tblastn with e-value of 10^{-2}. Blast hits which were mapping to the original reference protein were filtered. The blast result was further treated by solar and low-quality records (solar score < 25 and alignment rate < 0.25) were filtered. We extended the mapped region of each reference protein by 2000 bp upstream and downstream. Then GeneWsie was used to predict gene structure. The predicted gene fragments were filtered if they are overlapped with any genes in our millet gene set. Finally, pseudogenes were defined as parent-derived models with disablements (premature stop codon or frame shift) and

Nature Biotechnology: doi:10.1038/nbt.2195

covering at least 70% of the parent protein and at least 60% identity. In total, we detected 1,367 pseudogenes (**Supplementary Table 17**) in the millet genome.

Genomic variations between "Zhang gu" and "A2"

We identified the genomic variations, including SNPs, indels and SVs in the "A2" accession comparing to "Zhang gu" reference. In total, we identified 542,322 SNPs, 33,587 indels and 10,839 SVs. The average length of indels is 1.7 bp, and the frequency of indels with different length shows that the longer indels were rare (**Supplementary Fig. 13**). For SVs, we identified mainly four kinds of SVs, including insertions, deletions, inversions and the complex SVs. The frequency of those different categories of SVs reflected different detection power of our method, by which more deletions can be identified and false negative of insertions was higher. For SVs which have polymorphisms in length, the average length is 2827.7 bp. And the frequency pattern of different SVs in length (**Supplementary Fig. 14**) showed that the SVs in genic regions were shorter than SVs in intergenic regions, and the intergenic SVs were more abundant with the length of ~500 bp.

For SNPs, we checked the substitution between different bases (**Supplementary Fig. 15**). We found that of those SNPs, the substitution between purine to pyrimidine (transversions) was less common than that within purine and pyrimidine, which is consistent with previous study[5]. We then annotated the SNPs according to the gene annotation. In total, 119,773 (22%) SNPs were located in the genic regions and 422,549 (78%) SNPs were located in the coding sequences. The variation rate in genic and coding regions was less than that of the intergenic region and the whole genome average. For the 48,228 SNPs located in the coding regions, 20,720 SNPs were synonymous SNPs which would not result in amino acid changes and 27,508 SNPs were nonsynonymous SNPs which would change the amino acids.

Drought resistance related genomic features

Foxtail millet is a drought tolerant species, thus it is one of the most widely

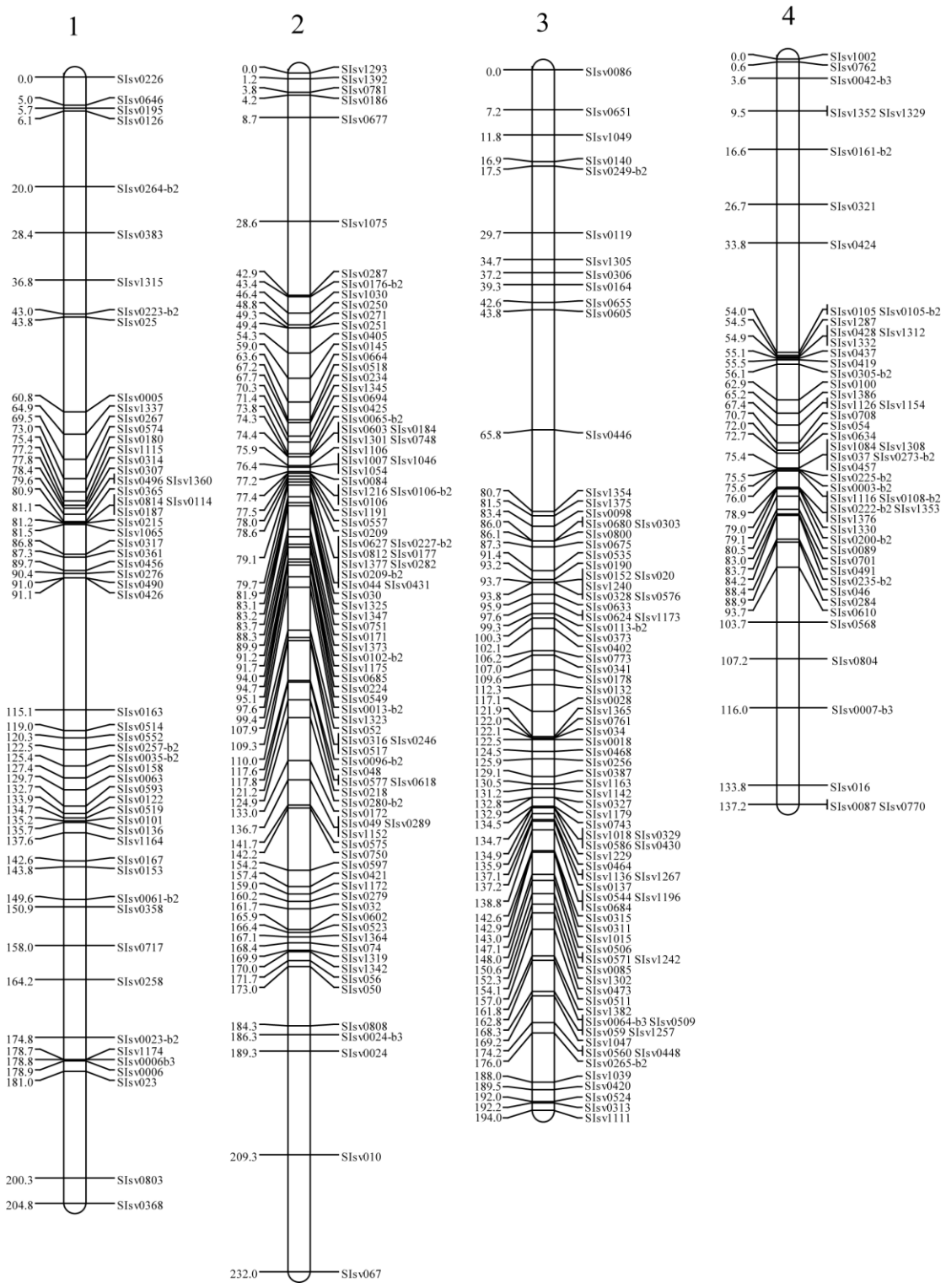
Nature Biotechnology: doi:10.1038/nbt.2195

grown crops in semi-arid areas. Many factors are involved in drought tolerance, including recruitment of genes, functional divergence, and gene expression variations controlled by key genetic regulators.

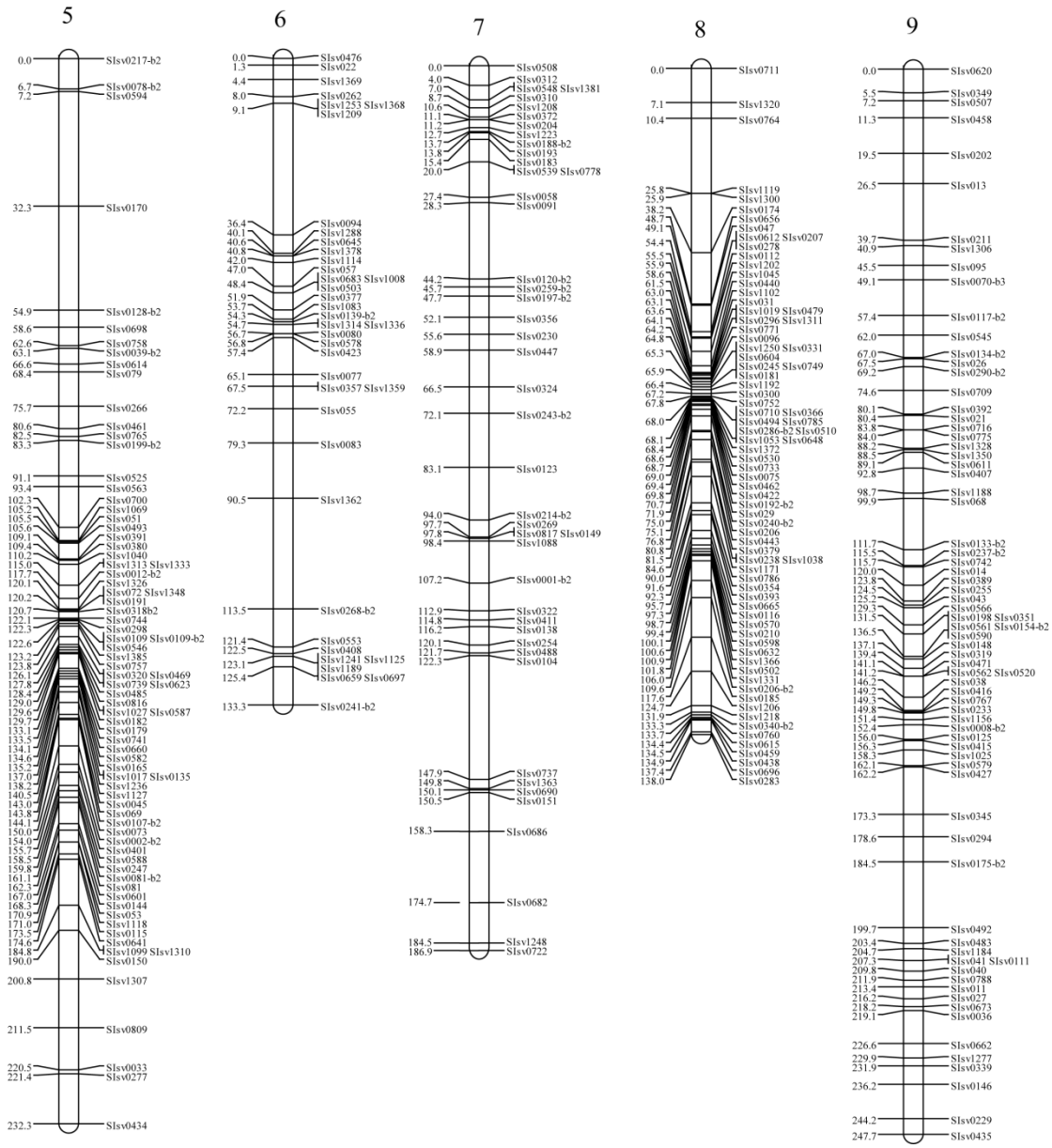
To investigate gene expression factors, we looked for miRNAs in foxtail millet and identified seven homologues of miRNA 169g, a drought inducible miRNA gene that is upregulated by the major upstream cis-element with DRE (dehydration-responsive element)/CRT (C-RepeaT) box[6]. Previous reports have identified miRNA 169g genes in rice (one), sorghum (five) and arabidopsis (one) [6-8]. Comparative genomic analysis of the well-established targets (the family of nuclear factor Y (NF-Y) B transcription factors) of the miRNA 169g subfamily in sorghum, maize, and *Arabidopsis* [8,9] in the foxtail millet genome showed that six members of this family were dispersed throughout the millet genome. NF-YB transcription factors are related with improved performance of *Arabidopsis* under drought conditions through the regulation of constitutive expression of specific protein-coding-genes[9].

Nature Biotechnology: doi:10.1038/nbt.2195

Supplementary Figures

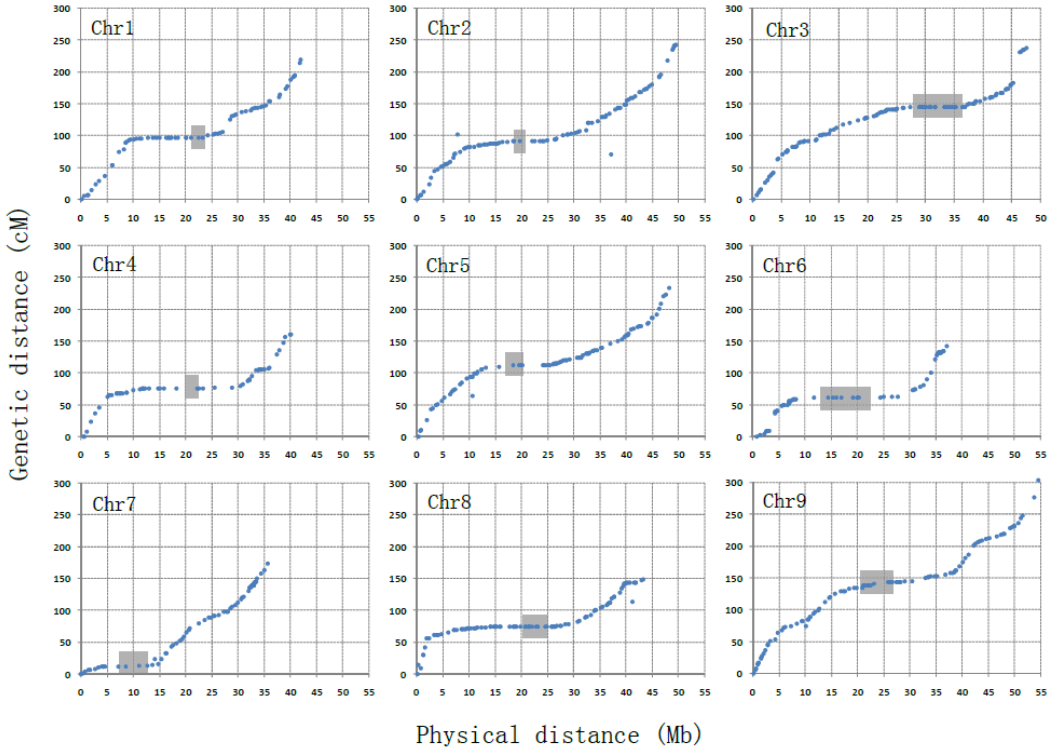


Nature Biotechnology: doi:10.1038/nbt.2195



Supplementary Figure 1 Genetic map of foxtail millet The genetic map was constructed by 751 markers (SVs and SNPs) in F2 individuals of "Zhang gu" and "A2". Join Map[10] was used to construct the genetic map. And all the primers of these markers and the genotypes of these markers can be found in **Supplementary Table 3**. The genetic map was 1,865 centimorgan (cM) in total.

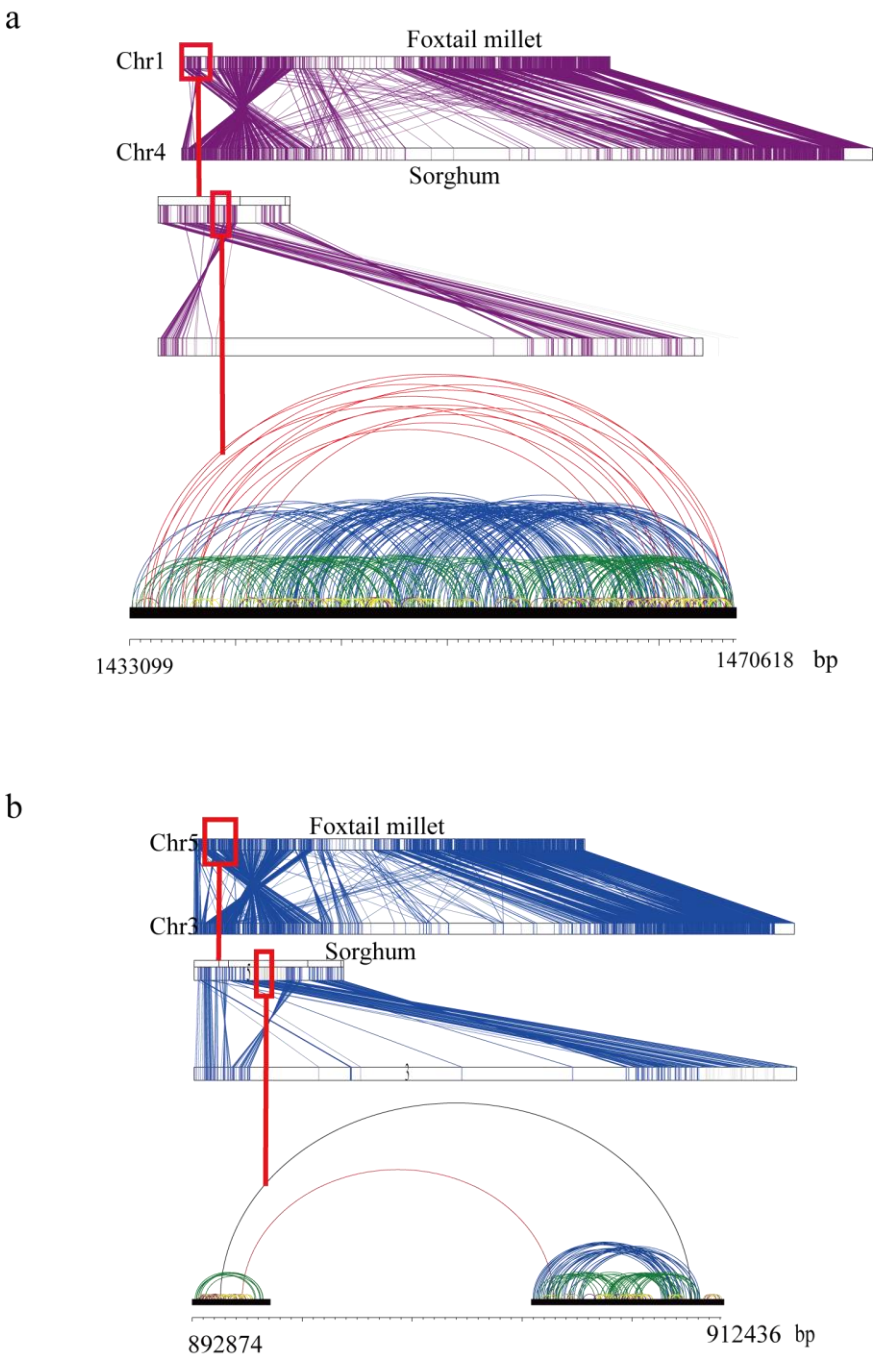
Nature Biotechnology: doi:10.1038/nbt.2195



Supplementary Figure 2 Genetic distance *vs.* physical distance Genetic position of the 751 genetic markers was plotted against the corresponding physical position. Comparing the centromere related repeat found in sorghum[8] to the foxtail millet genome, we have identified the 155 bp repeat elements which were abundant in regions with low ratio of genetic distance to physical distance. Those regions were shown in the grey box in the figure. And the sequence is as follows,

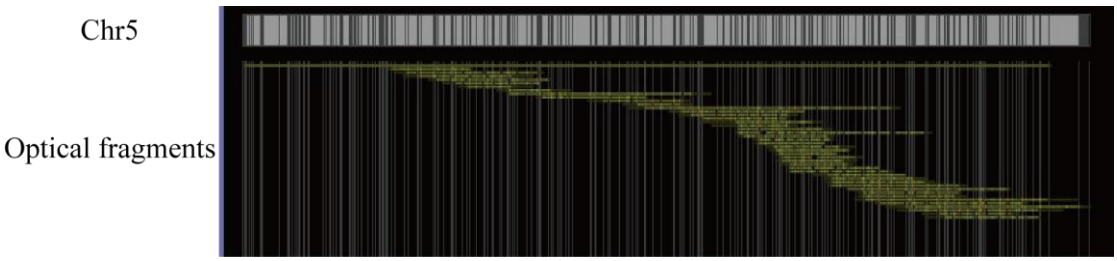
"TTTTGCATCATTTTTCGTGCCGGAACCGAATGCCCAAAAACACTCCCAAA CATGTTCTAGTGCATATTTAGGAAGATTGCATGCGTTCGTTGTGAAAAAATC CACCCGAGTTTCGCTACCCGGCAATAGTGCATTCGGGTGCGGAAATGCACC CG".

Nature Biotechnology: doi:10.1038/nbt.2195

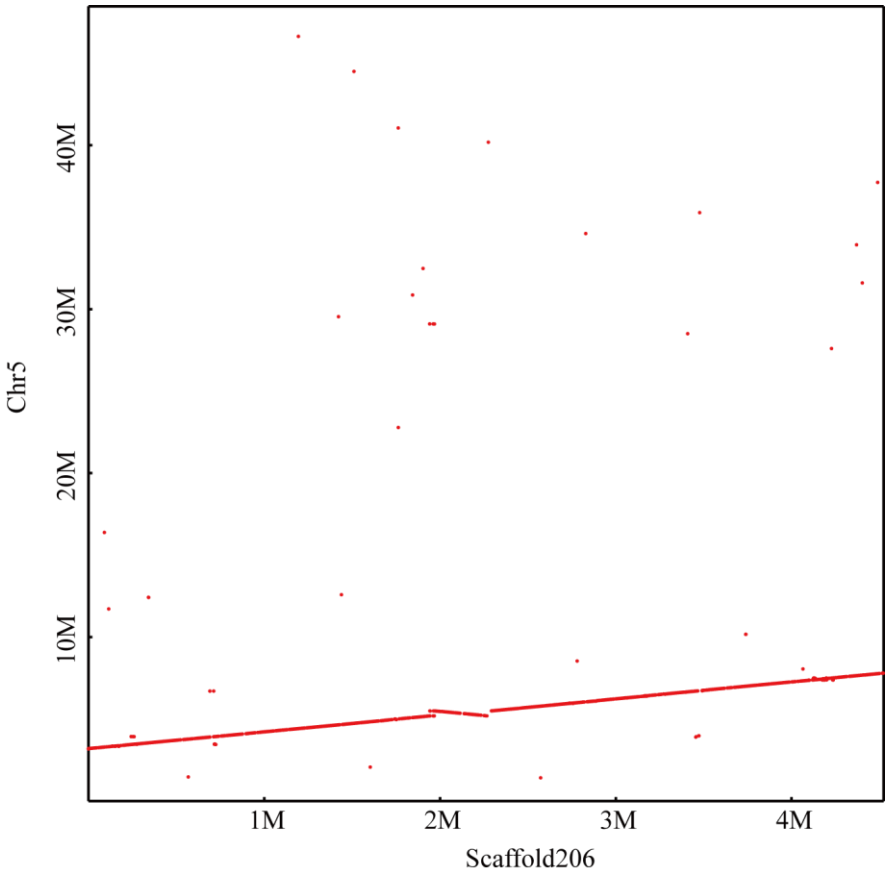


Supplementary Figure 3 Large scale rearrangements The rearrangements were identified between foxtail millet and sorghum using synteny identification. And then the mapping result of all the reads back to the reference genome was used to provide the information of support for those rearrangements. Of all the 1,937 rearrangements, only 7 were not well supported by pair end mapped reads. **a)** One rearrangement well supported by pair end mapped reads supported; **b)** one rearrangement without strong support from the pair end mapped reads.

Nature Biotechnology: doi:10.1038/nbt.2195

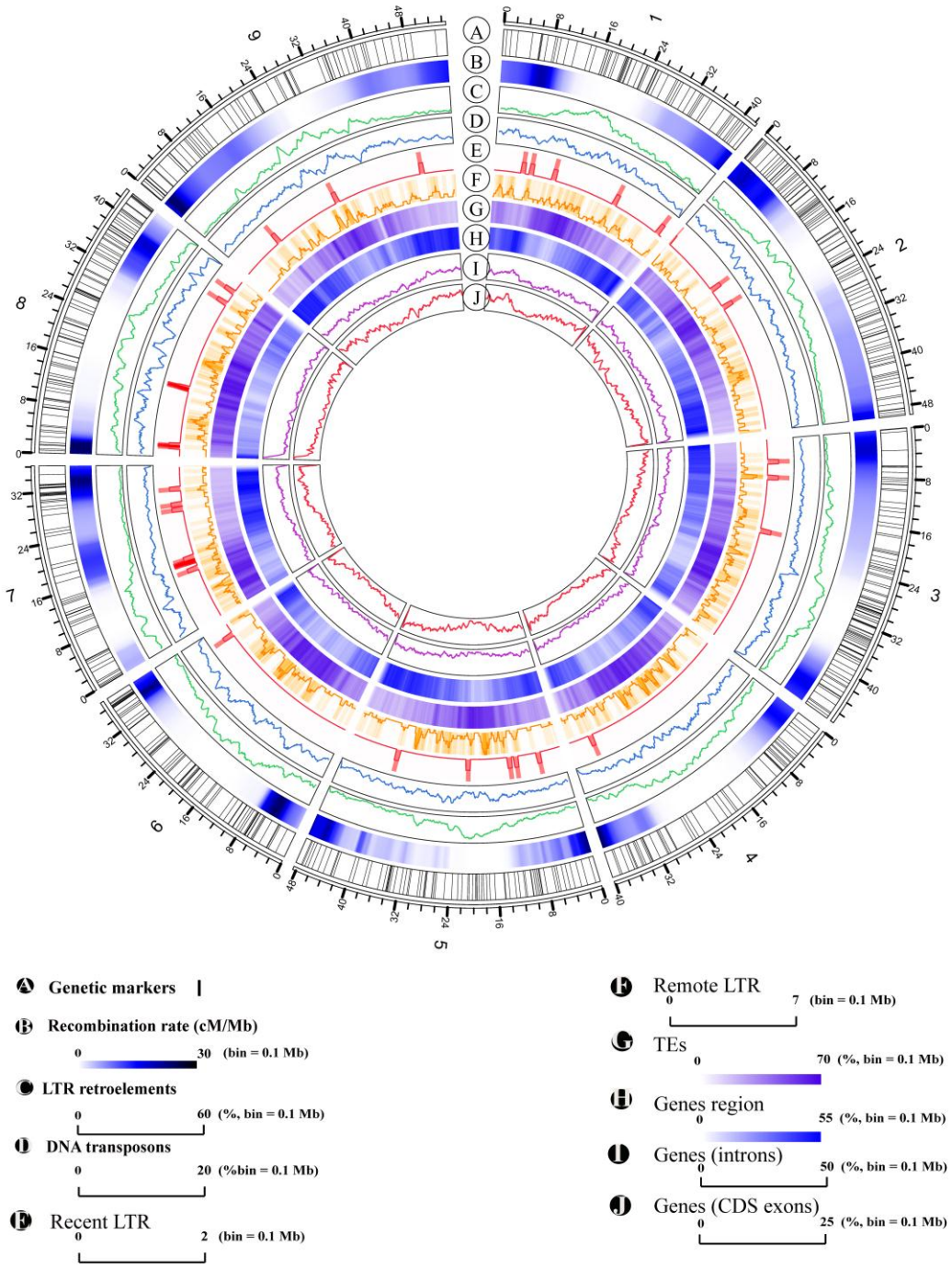


Supplementary Figure 4 Mapping optical fragments to the genome assembly The optical reads were generated by restriction enzyme and the optical machine (OpGen). Those reads were mapped back to the genome assembly. The mapped reads were indicated as the yellow lines. The reference was indicated as the above parallel. And the grey vertical lines indicate the restriction enzyme recognition sites on the reference genome determined by the sequence.



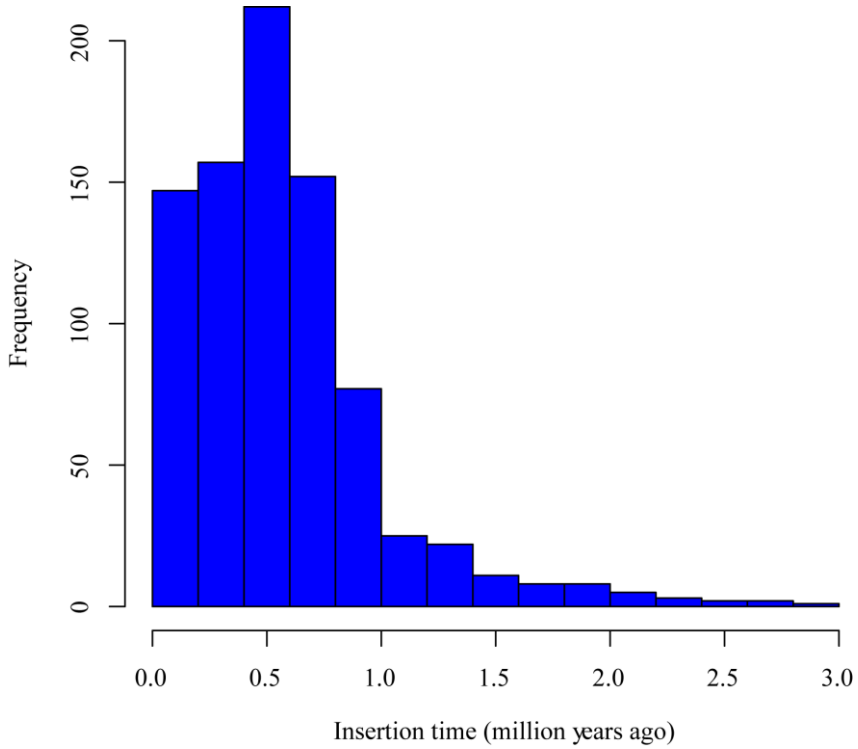
Supplementary Figure 5 Comparison of assemblies We compared the 20 longest scaffolds assembled by ALLPATHS-LG to our final chromosome assembly by Mummer. One scaffold (x-axis) was aligned against Chr5 of our genome assembly (y-axis).

Nature Biotechnology: doi:10.1038/nbt.2195



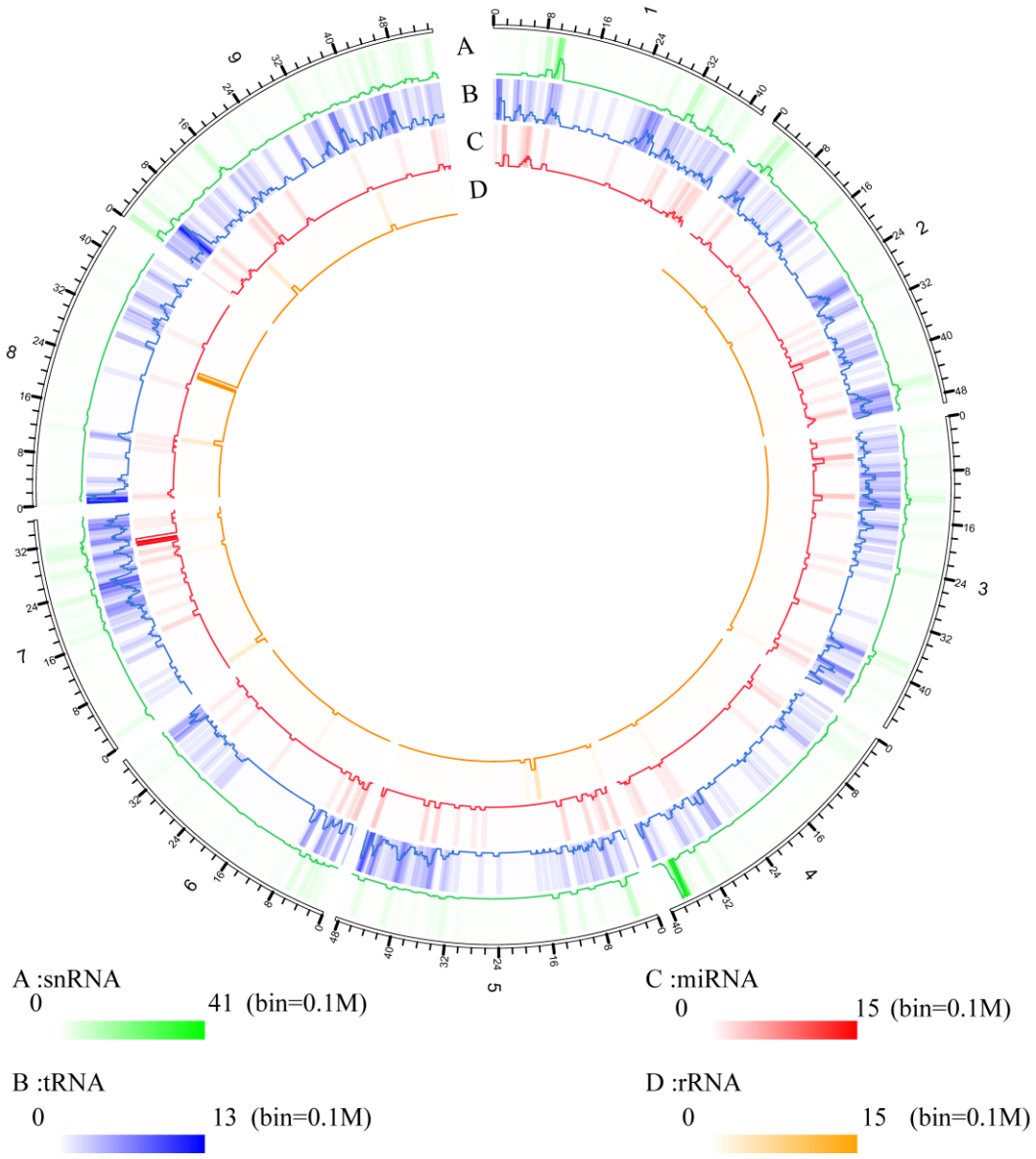
Supplementary Figure 6 Distribution of TEs across the genome Distribution of TEs (circle G) across the genome were compared with the distribution of genes (circle F) and the recombination rate (circle B). Regions with high ratio of TEs had low ratio of gene density and low ratio of recombination rate. Recent TEs (circle E) were distributed more randomly than ancient TEs (circle F).

Nature Biotechnology: doi:10.1038/nbt.2195



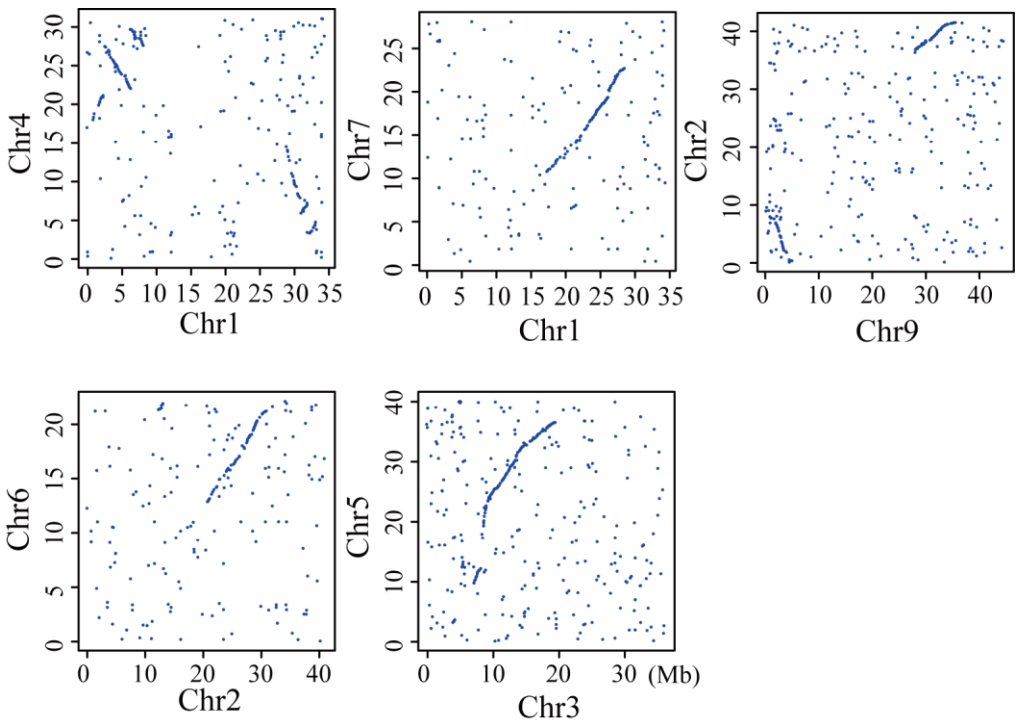
Supplementary Figure 7 Distribution of insertion time of LTRs Complete LTRs were identified in foxtail millet genome and sequence divergence of those LTRs was determined and insertion time of LTRs was then determined.

Nature Biotechnology: doi:10.1038/nbt.2195

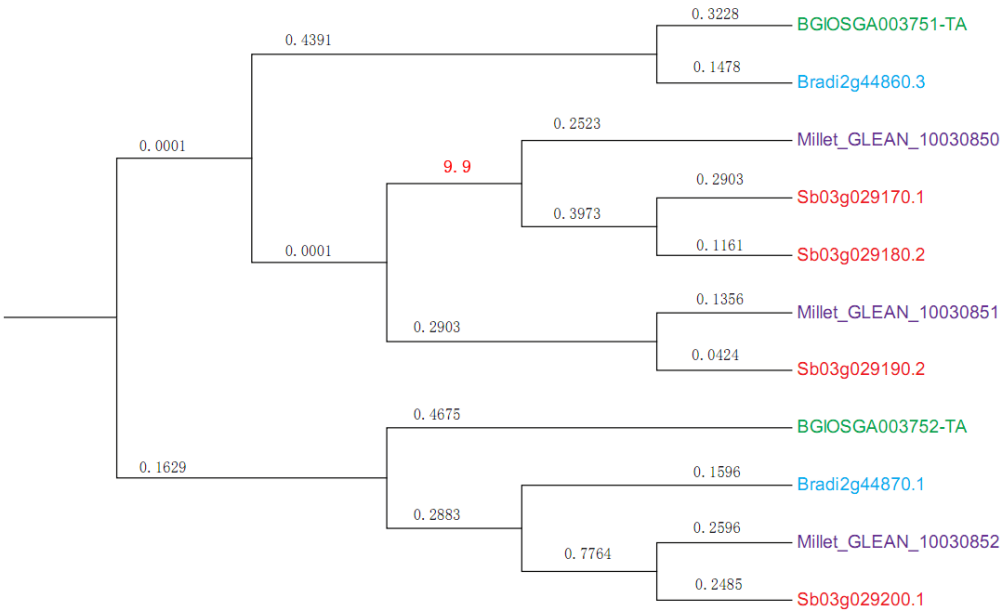


Supplementary Figure 8 Distribution of non-coding RNAs The density of non-coding RNAs were calculated in every 0.1 Mb bin of the genome. Biased distribution of non-coding RNAs in different chromosomes can be observed.

Nature Biotechnology: doi:10.1038/nbt.2195



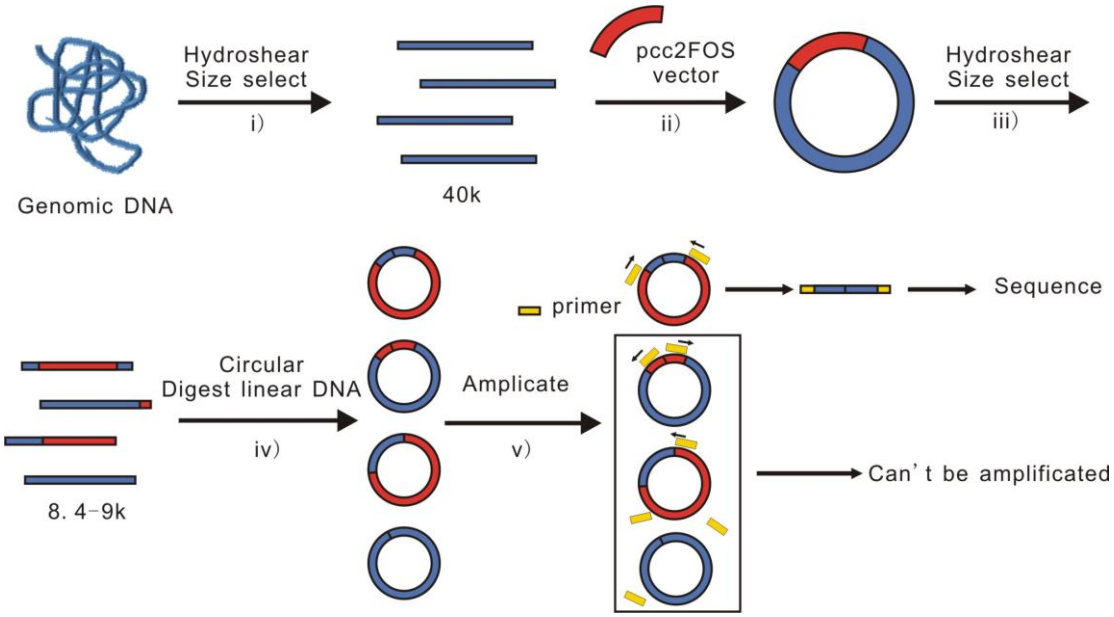
Supplementary Figure 9 Synteny blocks within foxtail millet The gene set of foxtail millet was compare to itself thus the paralogous genes in the foxtail millet genome were identified. The long synteny blocks were found, between Chr1 & Chr4, Chr1 & Chr7, Chr2 & Chr9, Chr2 & Chr6 and Chr3 & Chr5, which were resulted from the ancient whole genome duplication.



Supplementary Figure 10 The phylogeny of *CAβ* genes Phylogenetic relationship of carbonic anhydrate genes (*CAH*) from selected C4 plants (sorghum, foxtail millet)

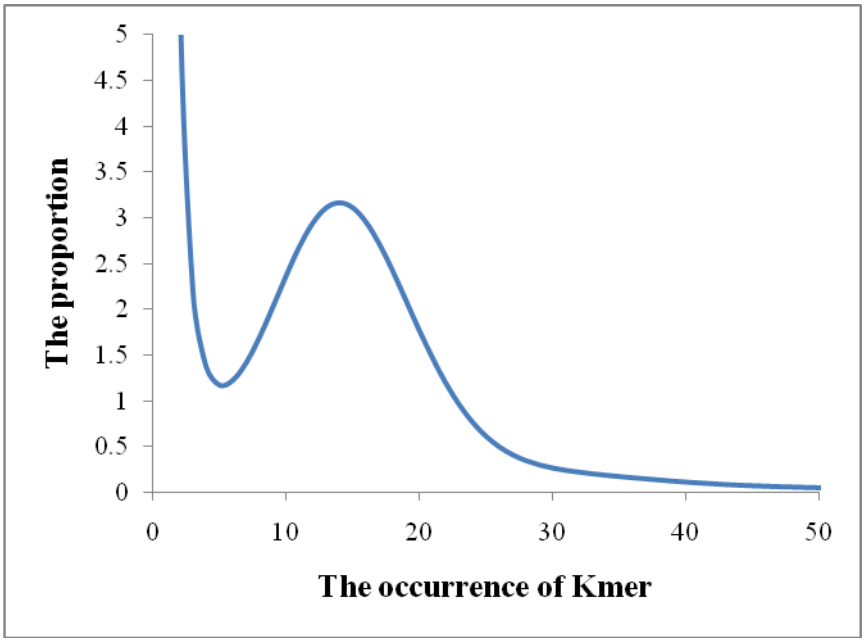
Nature Biotechnology: doi:10.1038/nbt.2195

and selected C3 plants (rice, *Brachypodium*) were shown. Values indicated on the branch were the ratio of nonsynonymous substitution to synonymous substitution (Ka/Ks). The tree was constructed by Bayesian method (implemented in mrBayes). The values of Ka/Ks were calculated using PAML[11]. The *Ft_CA1* gene, Millet_GLEAN_10030850, along with other two copies from sorghum, was all under substantially selection with high nonsynonymous to synonymous ratio (in red).

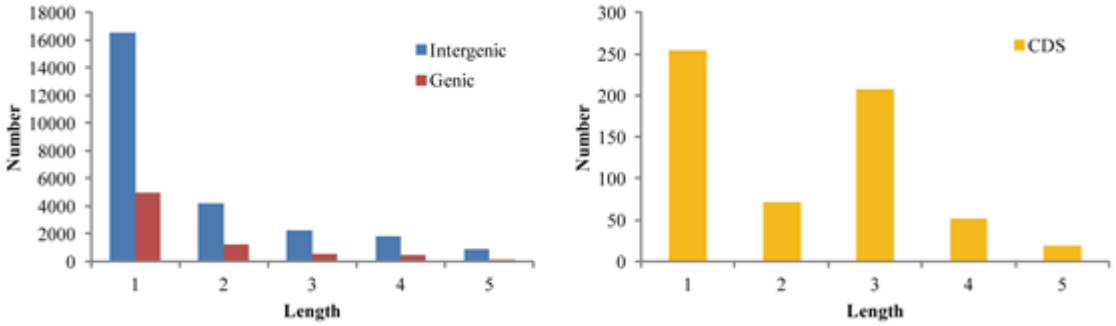


Supplementary Figure 11 Procedure of fosmid library construction i) Genomic DNA was randomly sheared to yield DNA fragments and size selected to obtain DNA fragments of approximately 40 kb in length. ii) DNA fragments were ligated to pCC2FOS Vector. iii) Shearing the plasmid DNA to generate fragments of 8.4-9Kb, we intended to get the fragments with both ends of inserts flanking fosmid vector. iv) Fragment circularization of digested linear DNA. v) Using primers which were designed according to flanking regions of the fosmid vector to amplify the end region by inverse PCR and then sequencing.

Nature Biotechnology: doi:10.1038/nbt.2195

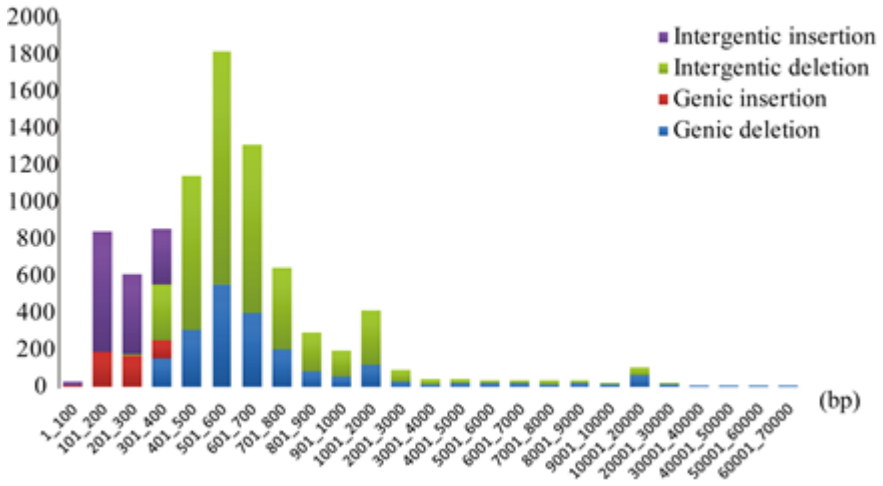


Supplementary Figure 12 k-mer analysis to estimate the genome size The figure shows frequency of 17-mers which are 17 bp sequences from the reads (after filtering) of short insert size libraries. We identified 6,793,533,642 k-mers using 17.7X data. The genome size can be estimated by (total K-mer number)/(the volume peak),which was thus estimated to be 485.3M.

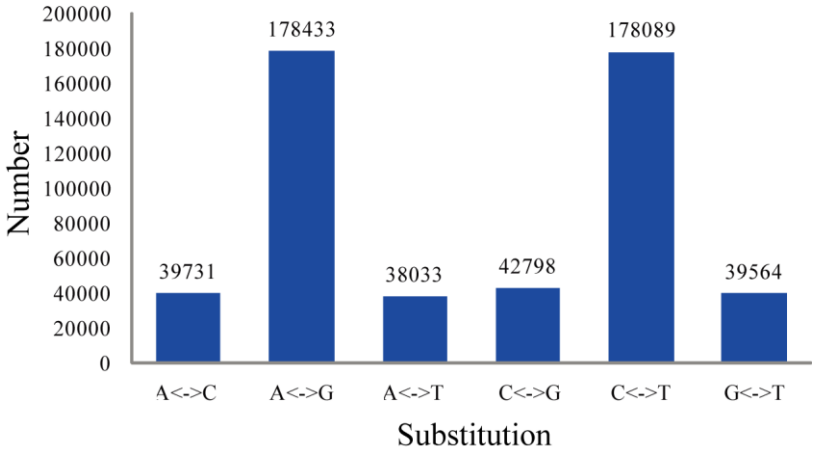


Supplementary Figure 13 Length distribution of indels Small indels were identified between "Zhang gu" and "A2". The length distributions of all the indels in different regions were shown.

Nature Biotechnology: doi:10.1038/nbt.2195



Supplementary Figure 14 Length distribution of SVs Structure variations were identified between "Zhang gu" and "A2". The length distribution of intergenetic SVs and SVs located within or nearby (within 1 kb) gene regions (genic insertion & deletion).



Supplementary Figure 15 Substitution of bases SNPs between "Zhang gu" and "A2" were shown according to changes of bases.

Nature Biotechnology: doi:10.1038/nbt.2195

Supplementary Tables

Supplementary Table 1 Libraries construction and sequencing DNA libraries with different insert sizes were constructed and sequenced. In total, 63.5 Gb raw data has been generated and the sequencing depth is about 127X. And after data filtering, about 80X clean data were used in the genome assembly.

Pair-end libraries	Insert Size	Total Data(G)	Reads Length	Sequence coverage (X)
Filtered Reads	170bp	11.2	75	22.5
	200bp	5.52	50	11.0
	500bp	4.47	50	9.0
	800bp	4.26	50	8.6
	2kb	5.66	44	11.3
	5kb	2.63	44	5.3
	10kb	5.60	44	11.2
	20kb	1.17	50	2.3
	40kb	0.089	40	0.2
Total		40.60		81.4

Supplementary Table 2 The assembly by SOAPdenovo The contig N50 was 25.4 kb, which means half (in length) of the contigs were longer than 25.4 kb. And the scaffold N50 is 1.0 Mb. The total assembly of scaffolds was 423 Mb in length, with 28 Mb Ns.

	Contig		Scaffold	
	Size(bp)	Number	Size(bp)	Number
N90	5,326	16,903	257,842	439
N80	10,497	11,748	465,544	313
N70	15,286	8,666	657,999	237
N60	20,058	6,419	865,534	181
N50	25,351	4,667	1,007,752	136
Longest	216,717		4,589,508	
Total Size	394,072,568		423,292,236	
Total Number(>100bp)	71,571		37,727	

Nature Biotechnology: doi:10.1038/nbt.2195

Total Number(>2kb)	23,451	1210
Length of Ns (Mb)	-	28

Supplementary Table 3 Markers in the genetic map 751 markers used to construct the genetic map and the primers of those markers were listed. The genotypes of the 480 individuals were also included.
(Included in a separate file)

Supplementary Table 4 Mapping reads to the genome assembly The mapping rates of reads from short insert size libraries were calculated for both single end mapped and pair end mapped.

Insert size	Statistic		
	Total pairs (PE)	Pair end (%)	Single end (%)
170bp	44487215	76.78	7.76
350bp	9628242	83.51	8.04
500bp	9788145	83.44	9.44
800bp	16469419	85.65	9.34

Supplementary Table 5 Large scale rearrangements in millet genome The genome of foxtail millet was compared to sorghum and the large scale rearrangements were identified.
(Included in a separate excel file)

Supplementary Table 6 Assembly by ALLPATHS-LG The contig N50 was 26.7 kb and the scaffold N50 was 2.2 Mb. The total length of scaffold was 412.8 Mb, which was shorter than that of SOAPdenovo.

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	5,781	14,490	363,012	215
N80	10,521	10,009	765,541	143
N70	15,576	7,253	1,173,476	99
N60	21,053	5,290	1,722,737	69
N50	26,708	3,794	2,232,506	49
Longest	206,753		10,810,433	
Total Length	355,458,898		412,844,823	
Total number>=100bp	29,355		4,633	

Nature Biotechnology: doi:10.1038/nbt.2195

Total number>=2000bp	21,789	2,430
Length of Ns (Mb)	-	57

Supplementary Table 7 Gene region coverage assessed by transcriptome The RNA sequencing data were assembled and aligned to the genome assembly. The proportion of RNA assembly aligned to the genome assembly was used to represent the gene region coverage.

Root							
Dataset	**Number**	**Total length**	**Covered by assembly (%)**	**with >90% sequence in one scaffold**		**with >50% sequence in one scaffold**	
				Number	**Percentage**	**Number**	**Percentage**
>200bp	61163	43759529	96.45	56014	91.58	60265	98.53
>500bp	27827	33295758	96.19	25359	91.139	27343	98.269
>1000bp	13591	23225397	95.82	12134	89.28	13293	97.81
Leaf							
Dataset	**Number**	**Total length**	**Covered by assembly (%)**	**with >90% sequence in one scaffold**		**with >50% sequence in one scaffold**	
				Number	**Percentage**	**Number**	**Percentage**
>200bp	46073	29649238	96.78	42616	92.50	45092	97.87
>500bp	19491	21456955	97.01	18264	93.70	19239	98.71
>1000bp	8675	13766019	96.63	8022	92.47	8528	98.31
Spica							
Dataset	**Number**	**Total length**	**Covered by assembly (%)**	**with >90% sequence in one scaffold**		**with >50% sequence in one scaffold**	
				Number	**Percentage**	**Number**	**Percentage**
>200bp	71940	50615396	96.62	66325	92.19	70936	98.60
>500bp	31707	38177066	96.36	29051	91.62	31206	98.42
>1000bp	15536	26703086	95.90	13928	89.65	15223	97.99
Stem							
Dataset	**Number**	**Total length**	**Covered by assembly (%)**	**with >90% sequence in one scaffold**		**with >50% sequence in one scaffold**	
				Number	**Percentage**	**Number**	**Percentage**
>200bp	65080	48085029	96.61	60117	92.37	64225	98.69
>500bp	29612	36991663	96.31	27084	91.46	29134	98.39
>1000bp	14894	26551269	95.86	13348	89.62	14586	97.93

Supplementary Table 8 Gene region coverage assessed by CEGMA The core eukaryotic genes (CEG) were mapped against the genome assembly and the

Nature Biotechnology: doi:10.1038/nbt.2195

completeness of those genes were determined. 99% of those highly conserved genes can be found in our genome assembly which reflected high coverage of the gene region. The CEGs were downloaded and different groups of CEGs were mapped to the foxtail millet genome to determine both the completeness and the copy number.

	CEGs number	Mapped Proteins	%Completeness	Total	Average copy number
Complete	**248**	**246**	**99.19**	**458**	**1.86**
Group 1	66	66	100	111	1.68
Group 2	56	54	96.43	91	1.69
Group 3	61	61	100	107	1.75
Group 4	65	65	100	149	2.29

Supplementary Table 9 Coverage of known genes 28 gene sequences (coding sequences) from NCBI nucleotide database of foxtail millet were downloaded and mapped against the genome assembly. Coverage of those genes was calculated. Only one gene (indicated in red) was not well covered in our genome assembly.

GI number	Length	Alignment length	Blocks	Mis-match
326369338	2856	2855	1	3
144583519	1428	1428	3	8
306008592	1119	1105	2	4
9965318	1134	1116	4	3
301087450	2938	2904	6	7
44663003	898	860	11	2
40354189	2886	2877	12	16
20257596	3214	3200	9	10
29123375	7271	7271	33	17
15558946	7630	7612	34	8
114848845	1260	1259	6	3
29123369	7446	7428	33	1
16904375	851	833	3	2
17221838	294	294	3	5
17221837	344	344	1	8
17221834	291	291	2	7
17221833	344	344	1	9
71044448	1274	1258	12	0
89029800	1390	1358	2	0
30351051	1454	1427	7	3
218158042	2673	2672	6	119
124263780	1061	1045	9	3

Nature Biotechnology: doi:10.1038/nbt.2195

124054094	245	37	1	1
118340263	1328	1309	13	10
22002427	2115	2085	16	7
56798257	2646	2596	3	18

Supplementary Table 10 Incomplete gene fragments in the assembly Incomplete gene fragments were identified by homolog searching of the plant proteins. In total, there are about 16,073 such fragments without complete gene structure.

Gene set	Total genes	Without start codon	Without stop codon	With pre-mature stop codons	Exon/ intron lost	Frame shift	Fragmental partial genes*
S.bicolor	63,701	48,291	44,873	15,700	18,972	18,198	9,575
A.thaliana	41,234	37,263	35,139	3,413	16,069	4,633	10,686
Z.mays	42,248	29,785	28,630	4,250	12,628	6,912	7,285
O.sativa	74,934	63,246	58,014	13,894	18,962	15,060	13,831

* Partial gene criteria: Coverage < 0.5, Length of fragments > 50 amino acid, with stop codons or frame shift..

Supplementary Table 11 Validation of incomplete gene fragments We randomly chose 89 gene fragments to be validated. We designed primers for those fragments, according to the positions which caused the incompleteness of the gene structure. 82 of those fragments can be amplified by those markers and sequenced. And 81 of them were validated by Sanger sequencing.

(**Included in a separate excel file**)

Supplementary Table 12 Repeat elements in the foxtail millet genome Repeat elements were identified by different methods and then combined into the final repeat annotation. In total, 46% of the foxtail millet genome were repeat elements.

	Repbase TEs		TE protiens		*De novo*		Combined TEs	
	Length (Mb)	% in genome	Length (Mb)	% in genome	Length (Mb)	% in genome	Length (Mb)	% in genome
DNA	21.8	5.14	11.3	2.66	38.8	9.17	39.7	9.38
LINE	4.8	1.14	6.6	1.55	7.0	1.65	7.6	1.81
LTR	51.2	12.09	42.1	9.94	115.4	27.25	125.3	29.58
SINE	0.4	0.09	0	0	0.7	0.17	0.7	0.17
Other	0.2	0.05	0.5	0.11	1.1	0.25	0.5	0.11
Unknown	0.1	0.03	0.2	0.04	28.2	6.68	22.8*[a]	5.39

Nature Biotechnology: doi:10.1038/nbt.2195

Total	78.5	18.55	60.7	14.30	191.2	45.17	196.6	46.44

a. The unknown TEs were combined from different TE prediction methods. If some regions were predicted as unknown by some method while classified into categories by other methods, those regions were not combined into the final unknown TE, thus the total length is less than that predicted by *de novo* method.

Supplementary Table 13 Categories of DNA repeats in grasses The repeat elements in foxtail millet were categorized into super families and the proportion of those families were determined and compared to other grasses[8,12-14]. Superfamilies listed here which were not reported in the corresponding paper were presented as '-' and other superfamilies not listed here but reported in the corresponding paper were classed into 'other'.

Super-family		**rice**	**sorghum**	**maize**	***Brachypodium***	**foxtail millet**	**Total length (millet, Mb)**
		% of genome					
Class I: Retroelements	**LTR/*Copia***	3.85	5.18	23.7	4.86	7.18	30.4
	LTR/*Gypsy*	10.9	19.00	46.4	16.05	22.14	93.7
	LTR/other	3.43	30.25	4.5	0.48	0.26	1.1
	LINEs	1.12	0.04	1	1.94	1.81	7.6
	SINEs	0.06	0	0	0	0.17	0.7
	Total classI	19.35	54.52	75.6	23.33	31.56	133.6
Class II: DNA transposons	**CACTA**	2.69	4.69	3.2	2.18	4.71	20.0
	***h*AT**	0.38	0.02	1.1	0.24	0.59	2.5
	Tourist	3.26	0.94	1.0	0.18	1.35	5.7
	Helitron	-	0.81	2.2	0.18	0.23	1.0
	Stowaway	-	0.19	0.1	0.88	0.61	2.6
	DNA/other	6.63	0.69	1.0	1.12	1.89	8.0
	Total class II	12.96	7.46	8.6	4.77	9.38	39.7
Other		-	-	-	-	0.11*a	0.5
Unknown (unclassified)		1.8	0.12	-	-	5.39	22.8
Total TEs		28.81	62.0	84.2	28.1	46.44	196.6

a. Other repeat elements here were tandem repeats.

Supplementary Table 14 Gene prediction in the foxtail millet Three gene sets were predicted independently and then combined to the final gene set, with 38,801 protein coding genes.

Nature Biotechnology: doi:10.1038/nbt.2195

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	**Augustus**	47,367	2169.21	788.67	3.12	252.49	650.09
	Fgenesh	49,206	2567.45	1085.22	4.54	233.31	426.86
Homolog	***A.thaliana***	41,234	1906.95	766.71	3.34	229.32	486.57
	O.sativa	74,934	1608.23	751.08	2.76	272.60	488.35
	S.bicolor	63,701	2297.51	989.27	3.49	283.26	524.87
	Z.mays	42,248	1937.57	835.37	3.45	242.18	449.98
RNASeq		31,709	3347.05	1144.05	5.07	225.82	541.79
Glean		38,801	2522.24	1087.33	4.25	256.10	442.10

Supplementary Table 15 Gene models in grasses The statistics of gene structure in foxtail millet were compared to other grasses.

	Rice	Sorghum	Maize	*Brachypodium*	Millet
Genome assembly size (Mb)	372.1	738.5	2,061.0	271.9	423.4
Chromosome assembly (Mb)	372.1	729.2	2,046.3	271.1	400.1
Unanchored Sequence (Mb)	------	9.3	14.7	0.8	23.3
# Genes	28,236	34,496	32,541	25,532	38,801
# Exons	134,812	136,658	287,136	140,142	164,737
Mean exon per gene	4.77	4.94	8.82	5.49	4.25
Mean exon size (bp)	364	297	305	268	256
Mean intron size (bp)	440	444	517	391	442
Mean intergenic region (bp)	10,339	17,002	122,608	7,311	17,895

Supplementary Table 16 Functional annotation of foxtail millet genes The genes of foxtail millet were mapped against protein databases and homologs with known function in each database were then used to assign functions for the foxtail millet genes.

		Number	Percent (%)
Total		38,801	--
Annotated	**Swissprot**	21,343	55.01
	TrEMBL	30,377	78.29
	InterPro	23,582	60.78
	KEGG	15,574	40.14
	GO	16,282	41.96
	Total	30,579	78.81
Unannotated		8,220	21.19

Supplementary Table 17 Pseudogenes in foxtail millet We identified pseudogenes

Nature Biotechnology: doi:10.1038/nbt.2195

in the foxtail millet genome and different categories were assigned.

	Number	Length (aa)	Coverage (%)	Protein similarity (%)
Duplicated	526	289	91.18	72.73
Retrotransposed	9	136	77.91	81.67
Unknown single exon	121	165	90.28	75.85
Unknown, multiexon	711	198	90.64	70.54

Length, Coverage and Protein similarity here refer to mean value;

#Duplicated: Both parent genes and pseudogenes are multi-exon genes;

#Retrotransposed: Parent genes are multi-exon genes & pseudogenes are single-exon genes;

#Unknown singleexon: both parent genes and pseudogenes are single-exon genes;

#Unknown, multiexon: parent genes are single-exon genes & pseudogenes are multi-exon genes.

Supplementary Table 18 Non-coding RNAs in the foxtail millet Noncoding RNAs were identified in the foxtail millet genome. The miRNA identified here were precursors.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		159	121.60	19334	0.00457
tRNA		704	75.02	52816	0.01248
rRNA		99	188.91	18702	0.00442
	18S	37	336.43	12448	0.00294
	28S	24	109.79	2635	0.00062
	5.8S	8	117.88	943	0.00022
	5S	30	89.2	2676	0.00063
snRNA		382	114.10	43586	0.01030
	CD-box	262	100.23	26259	0.00620
	HACA-box	31	122.58	3800	0.00090
	splicing	89	151.99	13527	0.00320

Supplementary Table 19 Conserved collinear blocks in foxtail millet Foxtail millet was compared with other grass genomes to identify the conserved collinear blocks.

foxtail millet *vs.* *Brachypodium*					
Chr2	98566	9011045	Brachypodium_Chr1	58795634	51708758
Chr2	40084993	49194169	Brachypodium_Chr1	23349939	13392989
Chr3	8470909	27702215	Brachypodium_Chr2	12904745	28701393
Chr3	15018645	19751145	Brachypodium_Chr2	51972907	54858606
Chr6	22868741	36976853	Brachypodium_Chr3	37310109	44360595
Chr6	148031	8858693	Brachypodium_Chr3	11607683	16934347
Chr9	41207104	55865684	Brachypodium_Chr1	60075252	74574281

Nature Biotechnology: doi:10.1038/nbt.2195

Chr9	88170	14383599	Brachypodium_Chr1	385454	12716468
Chr3	38340174	47354194	Brachypodium_Chr4	7039075	77444
Chr2	22906003	39694307	Brachypodium_Chr4	32082335	42927856
Chr7	13364443	32238684	Brachypodium_Chr5	10200651	25932025
Chr4	32030314	39978310	Brachypodium_Chr1	33934096	24715788
Chr4	122622	9995865	Brachypodium_Chr1	50730259	42272500
Chr1	103617	10932847	Brachypodium_Chr3	8182238	664586
Chr1	21663868	42098090	Brachypodium_Chr3	44503870	59650562
Chr9	32367685	38357022	Brachypodium_Chr3	36925524	33527752
Chr9	14666902	23952066	Brachypodium_Chr3	33296247	26325869
Chr5	24959767	48310858	Brachypodium_Chr2	41116117	59195467
Chr5	12584890	17527785	Brachypodium_Chr2	10418407	11781525
Chr5	5547669	14775362	Brachypodium_Chr2	317496	5413910
			foxtail millet *vs.* sorghum		
Chr8	24705	13691618	sorghum_Chr5	37565	20211306
Chr8	26048606	43314852	sorghum_Chr5	41367117	62328788
Chr1	21214872	42108033	sorghum_Chr4	40018921	68003094
Chr1	43638	10934098	sorghum_Chr4	13948846	706284
Chr3	8205733	27911179	sorghum_Chr9	59621314	40066749
Chr4	113762	17516703	sorghum_Chr10	14732	19948917
Chr4	27315560	36656758	sorghum_Chr10	56063437	49461187
Chr4	20033033	29796676	sorghum_Chr10	37954544	49365530
Chr3	38178984	47355654	sorghum_Chr8	55442543	39069671
Chr2	91822	18032519	sorghum_Chr2	124884	23949699
Chr2	27100851	49397517	sorghum_Chr2	55603207	77914305
Chr7	15161161	31377079	sorghum_Chr6	32734809	59387232
Chr7	107623	12562179	sorghum_Chr6	23542	7757322
Chr5	381920	14828791	sorghum_Chr3	15840088	1472447
Chr5	22527476	48324415	sorghum_Chr3	44665944	74438865
Chr6	1518009	8983819	sorghum_Chr7	9662904	1266551
Chr6	22780937	37265284	sorghum_Chr7	48827308	64258044
Chr9	28615235	55870264	sorghum_Chr1	42380652	73833605
Chr9	85128	24155256	sorghum_Chr1	2233	30661681
			foxtail millet *vs.* rice		
Chr1	103625	10933823	Rice_Chr2	9641015	562162
Chr1	20225760	42098096	Rice_Chr2	14340105	35856661
Chr1	11754937	16595350	Rice_Chr2	10351977	12859424
Chr7	15182496	31376231	Rice_Chr4	18380743	33615329
Chr9	38813176	55856181	Rice_Chr3	19006733	45042
Chr9	164650	14225658	Rice_Chr3	36108867	20040964
Chr3	13218649	19751100	Rice_Chr1	31049628	38003693
Chr2	27144990	39836615	Rice_Chr9	10805912	21957154
Chr2	109200	12830496	Rice_Chr7	10939	14703319
Chr2	40008559	49371571	Rice_Chr7	16806568	29607379

Nature Biotechnology: doi:10.1038/nbt.2195

Chr5	38244268	43998212	Rice_Chr5	25751395	20541317
Chr5	5081593	14817518	Rice_Chr1	6113	6321064
Chr5	24610061	48294882	Rice_Chr1	20188613	43124522
Chr9	14499744	26026398	Rice_Chr10	19577272	10275906
Chr9	32314129	38367134	Rice_Chr10	19621790	22671589
Chr6	25183965	37222659	Rice_Chr8	18954089	28348734
Chr6	3307116	8983389	Rice_Chr8	5143799	696444
Chr3	2366339	7994500	Rice_Chr5	188957	8351327
Chr3	8205724	27911128	Rice_Chr5	29544189	14720996
Chr4	30141948	36459984	Rice_Chr6	26953359	21072163
Chr4	128024	9896113	Rice_Chr6	141707	7978452
Chr3	38199795	47354194	Rice_Chr12	27552477	20059991
Chr8	51124	13089565	Rice_Chr11	67344	8870779
Chr8	30250744	35526972	Rice_Chr11	18504537	23044510
		foxtail millet *vs.* maize			
Chr8	21874613	29279848	maize_Chr4	17063499	23394860
Chr3	8205729	27898407	maize_Chr6	169240247	135887098
Chr1	27604666	37839838	maize_Chr4	120423542	166286101
Chr1	52571	10835794	maize_Chr4	228420485	245848088
Chr9	32899115	55870266	maize_Chr9	108543077	151339970
Chr7	12831599	31350920	maize_Chr10	104696950	145354112
Chr7	2165792	9328345	maize_Chr10	99364631	101764483
Chr7	15813726	31376966	maize_Chr2	72988924	4879597
Chr6	1533611	8983881	maize_Chr4	25847156	38385976
Chr3	38199769	46376176	maize_Chr3	131880840	113230185
Chr9	205498	26534633	maize_Chr5	657629	39537269
Chr4	2374395	15052647	maize_Chr6	108141871	121070057
Chr4	27820891	36460848	maize_Chr6	106039818	94611096
Chr2	22638953	49396577	maize_Chr2	168868914	215758862
Chr9	31275075	55847478	maize_Chr1	109471914	649624
Chr9	85129	14034180	Chr1	299544547	244243027
Chr9	14083075	24087043	Chr1	221338437	243882273
Chr5	37184518	48339309	Chr8	173914938	155123622
Chr5	25156713	36617919	Chr8	134400061	154955336
Chr5	381980	14692548	Chr8	166244	25405488
Chr6	22831258	30362005	Chr1	219369226	207953967
Chr4	116109	9896166	Chr9	29986399	15323
Chr8	30356732	39838354	Chr2	220353194	227617176
Chr5	22560993	48312643	Chr3	227066483	140039225
Chr5	1005859	14828791	Chr3	56896773	3128779
Chr2	27100838	49395376	Chr7	92961151	169216033
Chr2	92029	8681556	Chr7	179247	22511868
Chr1	21214872	42098098	Chr5	158211593	216474264
Chr1	10986010	19190518	Chr5	134903657	125105598

Nature Biotechnology: doi:10.1038/nbt.2195

Supplementary Table 20 Specific gene families in foxtail millet Comparing the gene families in foxtail millet to those of other grasses, the foxtail millet specific gene families were identified.

(Included in a separate file)

Supplementary Table 21 Contracted gene families in foxtail millet Comparing the gene families in foxtail millet to those of other grasses, gene families contracted in gene numbers were identified.

(Included in a separate file)

Supplementary Table 22 Genes in the C4 pathway The genes in the C4 photosynthesis pathway were identified by homolog searching in five grass genomes. And the genes in foxtail millet were listed.

Genes in C4 pathway	foxtail millet	sorghum	maize	*Brachypodium*	rice
***CAH* (carbonic anhydrase)**	4	5	5	2	2
***MDH* (malate dehydrogenase)**	7	5	6	3	3
***ME* (malic enzyme)**	6	7	6	7	7
***PEPC* (phosphoenolpyruvate carboxylase)**	6	6	6	6	6
***PPCK* (phosphoenolpyruvate carboxylase kinase)**	22	20	27	21	22
***PPDK* (pyruvate orthophosphate dikinase)**	3	2	2	1	2

	foxtail millet
CAH	Millet_GLEAN_10016964, Millet_GLEAN_10030852, Millet_GLEAN_10030850, Millet_GLEAN_10030851
MDH	Millet_GLEAN_10010264, Millet_GLEAN_10016287, Millet_GLEAN_10016288, Millet_GLEAN_10021967, Millet_GLEAN_10024194
ME	Millet_GLEAN_10002817, Millet_GLEAN_10008416, Millet_GLEAN_10020984, Millet_GLEAN_10022901, Millet_GLEAN_10026592, Millet_GLEAN_10032247
PEPC	Millet_GLEAN_10032141, Millet_GLEAN_10033023, Millet_GLEAN_10033129, Millet_GLEAN_10034966, Millet_GLEAN_10036278, Millet_GLEAN_10038607
PPCK	Millet_GLEAN_10002380, Millet_GLEAN_10002440, Millet_GLEAN_10002501, Millet_GLEAN_10003134, Millet_GLEAN_10004736, Millet_GLEAN_10005655, Millet_GLEAN_10009871, Millet_GLEAN_10011980, Millet_GLEAN_10013231, Millet_GLEAN_10013926, Millet_GLEAN_10016210, Millet_GLEAN_10020204, Millet_GLEAN_10020778, Millet_GLEAN_10023173, Millet_GLEAN_10023945, Millet_GLEAN_10025799, Millet_GLEAN_10026322, Millet_GLEAN_10026780, Millet_GLEAN_10028425, Millet_GLEAN_10029896, Millet_GLEAN_10031583, Millet_GLEAN_10038142

Nature Biotechnology: doi:10.1038/nbt.2195

PPDK	Millet_GLEAN_10022009, Millet_GLEAN_10011414, Millet_GLEAN_10022010

Supplementary Table 23 Phenotype of sethoxydim resistance The phenotypes of sethoxydim resistance of the F2 individuals were determined by growing the seeds of those individuals in the sethoxydim medium.

(Included in a separate file)

Supplementary Table 24 Markers used in fine mapping Five SNP markers in between the two SV markers were further used in the fine mapping of the sethoxydim resistance trait.

Marker	Chromosome	Start position	Forward primer	Reverse primer
SIsv0367	7	8637840	CGGTCATCACGTGCTCTATC	ATCTGCAGCCATCCTCTAGT
SNP1	7	9009560	CACATCCGAGTAGACTGGTC	CGAGCTCTGTGCCAGCTCGT
SNP2	7	9705710	ACTGCCTGCTATGCCACAAC	CAACCCTAGGATGCGTCCTT
SNP5	7	9806160	CATTCATAGACGGACCATCC	TTGTGAACCAAGACCTACTT
SNP3	7	10121440	CGCGAAGTATAACGTCAGAT	CACATGCTAAGGTACCGTTC
SNP4	7	10568320	CACAACACTTAGCGCATTTC	GATTTGTGCTAGTGTAGCCT
SIsv1223	7	11045880	TCAGTTTGCAGACCTGGACT	CAGAAGGAACCAAAGGCAGT

Supplementary Table 25 Raw data information of foxtail millet, SRA048234

Run Accession	Run Name	Reads Length(bp)	Insert Size(bp)
SRR424232.1	090502_I361_FC3044EAAXX_L8_SETzveDAADLAAPE	44	368
SRR424233.1	090613_I652_FC3130BAAXX_L2_SETzveDAADLAAPE	44	368
SRR424234.1	090902_I646_FC42KB3AAXX_L3_SETzveDAADLAAPE	44	368
SRR424235.1	090426_I58_FC301U9AAXX_L8_SETzveDABDIAAPE	75	352
SRR424236.1	090509_I638_FC312WLAAXX_L2_SETzveDABDIAAPE	44	352
SRR424237.1	090602_I331_FC303EFAAXX_L7_SETzveDABDIAAPE	75	352
SRR424238.1	090502_I354_FC304LJAAXX_L3_SETzveDABDIBAPE	44	388
SRR424239.1	090520_I637_FC303E2AAXX_L8_SETzveDABDIBAPE	75	388
SRR424240.1	090602_I331_FC303EFAAXX_L8_SETzveDABDIBAPE	75	388
SRR424241.1	090710_I77_FC42H5NAAXX_L8_SETzveDABDICAPE	44	503

Nature Biotechnology: doi:10.1038/nbt.2195

SRR424242.1	090714_I333_FC42HYFAAXX_L8_SETzveDABDICAPE	75	503
SRR424243.1	090719_I352_FC427DTAAXX_L2_SETzveDABDICAPE	75	503
SRR424244.1	090502_I361_FC3044EAAXX_L7_SETzveDABDWAAPE	44	2000
SRR424245.1	090519_I638_FC305BLAAXX_L1_SETzveDABDWAAPE	44	2000
SRR424246.1	090524_I361_FC3063GAAXX_L1_SETzveDABDWAAPE	44	2000
SRR424247.1	090831_I649_FC42K79AAXX_L8_SETzveDACDWAAPE	44	2000
SRR424248.1	091223_I637_FC61B3MAAXX_L2_SETzveDAFDLAAPE	44	5000
SRR424249.1	091111_I330_FC42T2CAAXX_L4_SETzveDAFDMAAPE	75	712
SRR424250.1	091111_I330_FC42T2CAAXX_L5_SETzveDAFDMAAPE	75	712
SRR424251.1	090618_I82_FC427J1AAXX_L5_SETzveDBDLAAPE	44	10000
SRR424252.1	090728_I77_FC42G66AAXX_L6_SETzveDBDLAAPE	44	10000
SRR424253.1	090930_I358_FC426NHAAXX_L4_SETzveDAFDBCPE	75	235
SRR424254.1	090930_I358_FC426NHAAXX_L5_SETzveDAFDECPE	75	400
SRR424255.1	091013_I80_FC42K6EAAXX_L5_SETzveDBDLAAPE	44	10000
SRR424256.1	091029_I321_FC42T5RAAXX_L3_SETzveDAGDTAAPE	44	10000
SRR424257.1	091029_I321_FC42T5RAAXX_L2_SETzveDAGDTAAPE	44	10000
SRR424258.1	101012_I114_FC80C0YABXX_L1_SETbnlDAADFAAPEI-10	90	40000
SRR424259.1	100917_I145_FC801W4ABXX_L3_SETbnlDABDBAAPE	100	170
SRR424260.1	100920_I162_FC801W3ABXX_L3_SETbnlDAFDUAAPE	90	20000

Nature Biotechnology: doi:10.1038/nbt.2195

Reference

1 Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317, (2010).

2 Huang, S. *et al.* The genome of the cucumber, Cucumis sativus L. *Nat Genet* **41**, 1275-1281, (2009).

3 Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239, (1988).

4 Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223-227, (2011).

5 Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195, (2011).

6 Zhao, B. *et al.* Identification of drought-induced microRNAs in rice. *Biochemical and Biophysical Research Communications* **354**, 585-590, (2007).

7 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).

8 Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556, (2009).

9 Nelson, D. E. *et al.* Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc Natl Acad Sci U S A* **104**, 16450-16455, (2007).

10 Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *The Plant Journal* **3**, 739-744, (1993).

11 Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Research* **20**, 1165-1173, (2010).

12 The map-based sequence of the rice genome. *Nature* **436**, 793-800, (2005).

13 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115, (2009).

14 Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763-768, (2010).